TR - H - 121

Proceeding of the ATR Workshop on "A Biological Framework for Speech Perception and Production"

河原　英紀
Hideki Kawahara

1995. 1. 23

ATR人間情報通信研究所
〒619-02　京都府相楽郡精華町光台2-2　☎ 0774-95-1011

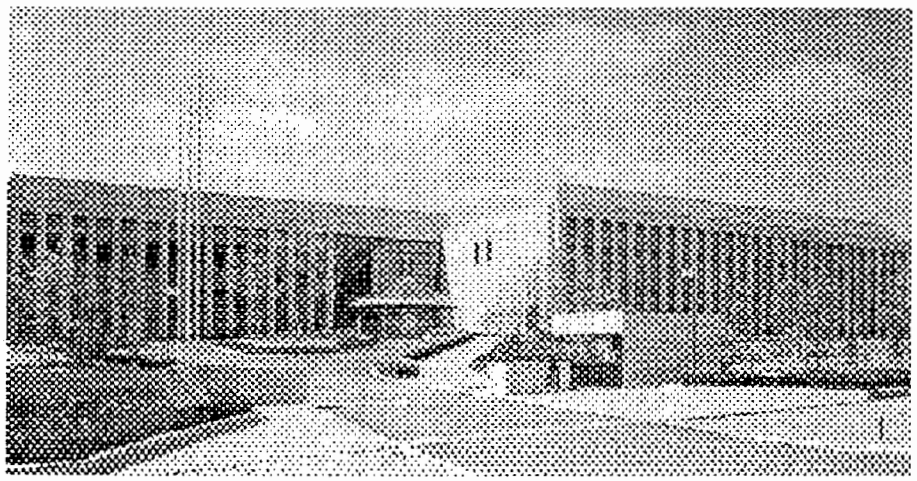
ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-774-95-1011
Facsimile: +81-774-95-1008

ATR

The ATR Workshop

A Biological Framework for Speech Perception and Production

September 16 and 17, 1994
ATR, Kyoto, Japan



ATR Human Information Processing
Research Laboratories

Organizer:

Yoh'ichi Tohkura
(ATR Human Information Processing Research Laboratories)

Chairman:

Hideki Kawahara
(ATR Human Information Processing Research Laboratories)

Co-chairman:

Roy Patterson
(MRC Applied Psychology Unit)

Speakers:

Robert P. Carlyon (MRC)
Martin Cooke (University of Sheffield)
Vincent L. Gracco (Haskins Laboratories)
William Morris Hartmann (Michigan State University)
Kiyoshi Honda (ATR)
Hideki Kawahara (ATR)
Stephen McAdams (CNRS)
Ray Meddis (Loughborough University)
Roy Patterson (MRC)
Malcolm Slaney (Interval Research Inc.)
Oded Ghitza (AT&T Bell Labs.)
Richard M. Stern (Carnegie Mellon University)
William A. Yost (Loyola Univ.)

Contents

Technical Session D

Commentary

Appendix

Opening Address for the ATR workshop on A Biological Framework for Speech Perception and Production

HABARA Kohei, Dr.

Executive Vice President, Research, ATR International
Board Chairman, Four ATR Consortia (Active R&Ds)
President, Two ATR Consortia (Research Fruits Managing R&Ds)

Greeting

Good morning, ladies and gentlemen. I would like to welcome you all to the second ATR workshop on speech perception and production. First, I would like to express our sincere appreciation for your participation in our workshop, especially to those of you who have traveled a long distance to attend.

I am sure that your excellent presentations, your active participation in discussions, and your exchange of new ideas, not only in the sessions but also during the breaks and at the party, will make this workshop a great success. With your help, the direct and indirect results of this workshop will serve for scientific advancement on our knowledge about speech communication. They will also contribute to improving our quality of life and the security of our world through future communications systems.

Philosophical changes

At this time, let me briefly talk about my personal opinion regarding the way we should think about future research in telecommunications technologies.

It has often been said that the objective of communication, especially telecommunications, is to overcome distance barriers. Telephony surely did this with voice communications for more than a century. So, what's next? The telecommunications industry is rapidly evolving from the plain old telephone service era of the last 100 years into a sophisticated multimedia era.

The ultimate form of communication is face-to-face interaction between human beings, utilizing all the five senses as we do in nature. In the actual telecommunications environment, however, human beings interface with networks or facilities. And human users have been forced to deal with the machine interface to achieve a face-to-face feeling, rather than the other way around. This could be likened to "Geocentrism" on the part of the service providers. The need to dial telephone numbers, and operate a computer via keyboard and mouse, are typical examples of the limitations of today's human-machine interface technology.

Regarding "Geocentrism," let me describe the contradictory terminologies such as "automatic " and "manual." In the early days of manual telephone switchboards, users placed calls by simply calling "Mr. A, please," or "the Sobashop at the corner, please." Thus, for these users, the telephoning process was automatic. Since the introduction of so-called "automatic switching systems," such as "step-by-step systems," users have had to dial the telephone by hand. For them, the "automatic" switching system is "manual."

"Automatic vending machines" are completely manual as far as users are concerned. The user has to find the coin slot, drop the necessary coins in the slot, and bend over before being able to pull a Coke and the change out of the machine. "What's automatic about this?"

Thus in order to produce a really human interface, we must first learn more about human functions. This will help us move the "conversion" from "Geocentrism" to "Heliocentrism," or in other words, from "manual" to "truly automatic" systems.

And, this is actually my philosophy in managing the research activities at ATR as well.

ATR's mission

ATR was established as a central institute of the Kansai Science City 8 and a half years ago. ATR's mission is to carry out cutting-edge research in telecommunications fundamentals, through four research and development consortia. Among them, the Human Information Processing Research Laboratory is the one studying sophisticated human functions. There are many difficult but exciting issues in this magnificent field. A "trans-disciplinary" approach, which I have been advocating, is required in addition to an "inter-disciplinary" approach.

Cultural issues

Research methodologies are strongly affected by the culture where the institution is located. ATR is located in the area of the old capitals of Japan, Nara and Kyoto, which have more than a thousand years of history. You can visit old temples, shrines and historical monuments. Researchers living in this unique mixture of tradition and innovation have the opportunity to think deeply about their role in terms of the whole of human history. In the long run, this will characterize our research and enable us to contribute to international societies in a unique way.

Here let me briefly touch upon Oriental culture. It could be said that Oriental methodology puts considerable emphasis on "synthesis" or "integration," as well as "analysis." This philosophy, I believe, will become more and more important when studying such highly sophisticated creatures as human beings. This implies the importance of the "harmonization" of Western and Oriental cultures.

Surrounding environment

Next I'd like to point out some coincidences, which are nevertheless symbolic for this workshop which is focusing on a "Biological Framework."

First, field trials of new optical communications systems and applications recently began in the houses and laboratories, within a stone's throw from ATR. The trials place emphasis on users' experiences with so-called multimedia. A Plenipotentiary Conference of the International Telecommunications Union of the United Nations, which will discuss future telecommunications frameworks, will also begin in Kyoto from next Monday. I will have the great honor of introducing one of our research results in the presence of Their Imperial Highnesses, the Crown Prince and Princess of Japan, on the opening day.

Second, Japan opened a new door to the world with Kansai International Airport a couple of weeks ago, which will greatly facilitate travel especially to the west and south directions.

Third, a series of ceremonies for the inauguration of Kansai Science City will start next Friday and last two months. The city is usually called "Kansai Science City." However, the precise English equivalent of its name should include the word "culture," which stems from the twelve hundred year history of this region. This is again an interesting coincidence for our workshop, where biological, in other words, natural, aspects and their relation to higher human functions are the main interest.

West meets east

The last and the most important point is that it is now time to bring Western and Oriental creativity together to solve the many complex and difficult problems which cannot be penetrated by a single paradigm. Investigating human speech communications is one such problem.

I am sure that the heated and lively discussions which are expected to take place in this workshop will initiate a number of trans-cultural and trans-disciplinary research collaborations to attack these long lasting questions.

Concluding remark

Finally, I would like to thank you all once again, in advance, for your active participation in our workshop.

Thank you.

Impact of Biological Aspects of Speech Perception and Production on Future Communication Systems

Hideki Kawahara

ATR Human Information Processing Research Labs.
2-2 Hikaridai, Seika-cho Soraku-gun
Kyoto 619-02, Japan
kawahara@hip.atr.co.jp

INTRODUCTION

> Text is not enough to represent speech.
> Speech is not enough to represent heart.
> *from a Chinese sutra of divination lore compiled in B.C. 800*

Rapid growth in information processing power and advances in device technologies will give humans a chance to re-integrate various communication channels that were split several thousands of years ago, probably due to limitations in the technologies which were available in those periods. This re-integration which will take place in the next century is not only a retention of the lost coherency, but also a revolution associated with enhancement.

The enhancement has resulted from the positive sides of the separation. The "Text" channel gained the power to break barriers of time, distance and distribution through Gutenberg's printing system. The "Speech" channel gained the power to break barriers of time and distance through Edison's recording system and Bell's telephone system. The "Vision" channel also gained similar power through the television system and video recorders. Current information technologies not only inherit these advantages, but also provide the means to represent all of these information modes in a uniform digital format. Uniform digital representation makes it potentially possible to convert from one channel to another. The re-integration and conversion of multi-modal information based on digital technologies will enable future communications systems to break barriers introduced by spatial and temporal separation, language, culture and disability. That will provide the ultimate prosthesis for communication difficulties usually caused by aging.

To make this possibility feasible, however, it is necessary first to break the hardest barrier, the representational barrier between humans and machines. Future communication systems have to implement algorithms to simulate the necessary transformations in order to be able to manipulate sensory information in a way that is meaningful to humans. Without proper representations, the degradation that is associated with media conversion will be perceptible and irritating. For example, current text–to–speech and speech–to–text conversion technologies are still far from satisfactory, especially under conditions which are common in everyday life. Of the heart–to–speech and speech–to–heart conversion, almost nothing is known.

Thinking about and investigating an integrated communications channel as a holistic entity based on a biological framework may be what is required to break this barrier. In the rest of this article, I will try to raise questions about how such advanced systems

should behave and how such systems should be designed. I will also provide a rough estimate of the computational power necessary to implement those functions.

BACKGROUND: Auditory Scene Analysis

"Auditory Scene Analysis" by Albert Bregman is an important step[1]. It proposes an ecological point of view in hearing research as a complement to the traditional psychophysical research. In a broader sense, auditory scene analysis (ASA) assumes that the primary function of our sensory systems is to extract important environmental information via multiple sensory modalities and to re-construct a coherent representation of the outside world: That is scene analysis. The representation has to be dynamic, meaning it has to represent not only "what is there", but also "what is happening".

As a part of this integrated multi-modal sensory system, the auditory system tries to solve this problem. However, re-constructing a dynamic representation of the outside world from one-dimensional (monaural) or two-dimensional (binaural) signals is an ill-posed problem. In other words, it is impossible to select a unique solution from a set of infinite numbers of possible answers, without using additional constraints. In his 1993 article[2], Bregman summarized some of these constraints into the following four heuristic rules.

1: Unrelated sounds seldom start or stop at exactly the same time.

2: Gradualness of change

   a) A single sound tends to change its properties smoothly and slowly.

   b) A sequence of sounds from the same source tends to change its properties slowly.

3: When a body vibrates with a repetitive period, its vibrations give rise to an acoustic pattern in which the frequency components are multiples of a common fundamental.

4: Many changes that take place in an acoustic event will affect all the components of the resulting sound in the same way and at the same time.

These rules are still not complete and not always true. In other words, they are probably approximately correct (PAC) rules. These rules, however, are crucially important for disambiguating ill-posed problems within a limited amount of time. Application of these rules can be seen as a dynamic process to make hypotheses and to test them based on available evidence. This hypothesis-and-test framework is formulated mathematically in vision research on 3D shape reconstruction from monocular 2D pictures[3]. There are many other computationaly equivalent formulations. Kawato's formulation is preferred as an algorithm level model in Marr's[4] terminology, because it can explain processing times found in psychological experiments and provides a relationship between neural structures and functions. In principle, a similar approach may be possible in auditory scene analysis. The major obstacle which hinders this approach is that it is not clear what has to be re-constructed for the auditory scene.

These active processes are found at every level in the auditory system, from otoacoustic emissions[5] to phonemic restoration. In the intermediate levels, we can find many auditory illusions[6] including auditory induction, the continuity illusion, the octave illusion

and so on. These may be implemented by bidirectional connections found in auditory pathways[7].

One important point to note here is that the key words in the rules, "same", "smooth", "slowly" and "same way", have to be fuzzy in advance. They have to be set based on the experience of exposure to some specific sound environment. These requirements coincide well with the fact that our auditory cortex is still plastic even in adults[8].

Another interesting fact about this list is that these rules sometimes conflict with speech properties. The speech signal consists of various types of discontinuities. Stop consonants and voiced to unvoiced transitions are good examples. Even within voiced sounds, dipthong, nasal to vowel and consonant to vowel transitions violate rule 4. At this point, a biological framework of speech perception and production provides additional constraints for segregating a speech signal as a single entity.

Auditory visual interaction in speech perception

The McGurk effect provides an interesting demonstration of auditory visual interactions in speech perception[9, 10]. It is striking to observe how easily visual signal can alter speech perception. However the McGurk effect itself is a kind of artifact. Auditory and visual interaction plays an important role in integrating cues to improve intelligibility under adverse conditions. It is important to note that the equivalent signal to noise ratio sometimes shows a 16dB improvement by introducing visual information. This is comparable to the maximum binaural masking level difference (BMLD)[11]. It is suggested that this integration of auditory and visual information takes place before categorical perception occurs[12].

This early integration may suggest that the internal representation of the outside world is inevitably multi-modal. A strong positive correlation (=0.91) between the visual and auditory acuity of various mammals also may support this hypothesis[13]. The fact that the correlation between auditory acuity and distance between both ears is weak (=0.57) for a set of similar mammals suggests that the reason for this coincidence between visual and auditory acuity may well be explained by ecological requirements and cannot be inferred from individual neurological or physiological data. This may be an example of when the use of a holistic point of view is important.

Knowledge about the developmental process may provide another example. Consider the interaction between a mother and her baby. They are using all available information channels to communicate. It is reported that auditory visual integration can be found even in very young babies[14]. This may indicate that multi-modal communication is the primary form of our communication.

Spoken language

Speech recognition is traditionally treated as the transformation of input sound space into discrete linguistic symbols. In the case of speech understanding systems, even though they have wider scope, they still deal with information which can be transcribed into text. It has to be pointed out again that this transcription into text is the splitting of an integrated entity into pieces based on a scheme introduced several thousand years ago. This may be a good time to think about portions which are removed in standard speech recognition systems.

Development

The vocal communication between a mother and her baby starts with motherese, characterized exaggerated prosodic patterns[15]. The special patterns in fundamental frequency have special effects on the baby's response. It should be noted that even a very young baby is able to mimic the essential characteristics of motherese.

Through these communications and exposure to a specific language environment, it is reported that infants develop their vowel system as perceptual magnets by 6 months of age[16]. Up to this moment, infants show a universal ability to discriminate phonological contrasts which will never appear in their mother tongue. This ability diminishes with time and almost vanishes by the age of 12 months[17].

For this kind of learning to be possible babies must have mechanisms to measure differences between their mother's voice and their own voice. This is not a straight-forward task, as differences in fundamental frequency and in vocal tract shape and size make it meaningless to directly compare physical parameters of speech sounds. Speech sounds have to be normalized in some fashion possibly through reference to the fundamental frequency. What kind of representation can have such property?

This developmental process seems to imply that our speech communication consists of a hierarchical structure constructed on a primary multi-modal representation subsumed by upper structures such as, vowel and consonant classifications, and prosodic information.

Second language acquisition

A cross-sectional study of 150 Japanese subjects who lived in America revealed that /r-l/ identification patterns are correlated with the initial ages when they started their stay in America. 90% of the subjects who started living in America before the age of 10 showed native-like dissemination, however, less than 10% of the subjects who started living in America after the age of 10 showed comparable performance[18, 19].

At first glance, this seems to suggest thét there is a critical period for second language acquisition. But that is not the case. With the combination of multiple talker and identification training with minimal word pairs, Japanese adults showed more than a 20% increase in correct response rate after a total of 20 hours training. This ability was generalized both to new talkers and new words, and retained for 6 months after training[20]. This provides evidence that the auditory system is plastic even in adults. This is a good news for second language learners, but of what use is it for the others? If the hierarchial structure suggested in the previous paragraph is valid, is a later structure more plastic?

Interactions between speech perception and production

It is believed that interactions exist between speech perception and production. There are several sources of evidence for this: Effects under delayed auditory feedback (DAF)[21] and the Lombard effect[22] are good examples. But the destructive effects of these conditions make it difficult to apply them to the investigation of these interactions under natural conditions.

A new measuring technique called TAF (transformed auditory feedback) has been developed to investigate the interaction between speech perception and production without disturbing natural speech production. TAF enables a quantitative analysis of interactions mediated by various parametric representations of speech. The first series of experiments

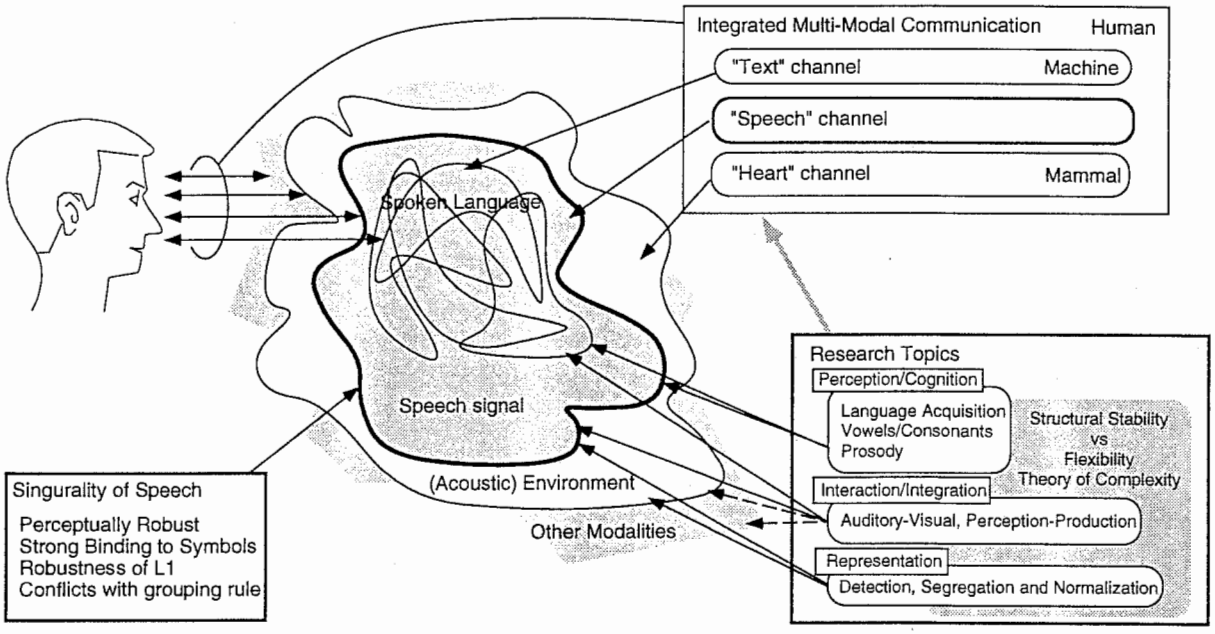


Figure 1: A schematic diagram of integrated multi modal communications and related research topics.

revealed that there are compensatory responses to fundamental frequency perturbations with about 150ms of latency. It was also confirmed that responses similar to those found under TAF conditions also exist under natural conditions[23]. Preliminary tests indicate that the compensatory response can be observed during intentional vibrato and also for whistling.

This suggests that the response found by TAF is an automatic response but also a postnatal function. It is possible to speculate the implication that pitch perception plays a key function in human speech communication, especially for information about emotional states and the transition of roles.

Communication system to mediate whole aspects

Figure 1 shows a schematic representation of re-integrated speech communication and necessary research topics for such communication to be mediated by future communications systems.

Computational power to simulate the auditory system

The recent rapid growth of computational power is driven by improvements in integrated circuit technology. Trends in microprocessor and mainframe CPU performance growth clearly indicate this[24]. The performance of microprocessor-based machines has increased at a rate of 100 times per decade. If we extrapolate this trend, GFLOPS machines will be at hand in the beginning of the 21st century.

Hans Moravec estimated the computational power to simulate the whole human brain function based on the temporal and spatial resolution of a retina[25]. He suggested the computational power of the whole brain to be 10^{13} operations per second. By extrapolating the trend of growth in computational power, it is expected that cutting edge

computers will have equivalent computational power around the year 2005, and personal computers will have the same power around the year 2030. This is the upper bound. Even if we made an underestimation by a factor of 100, these dates would advance only 10 years toward the future. Since the auditory system is a subsystem of the whole, real time simulation of a full-scale model will become possible well before these dates. Finer simulation of each cell level may need to increase the number of operations 10^4 to 10^5 times[26]. This will add another 20 years to the estimate.

On the other hand, if we can model at a functional level, a reduction in computation on the order of 100 will be possible. This means that if we can find a proper functional model of the whole auditory system, we will be able to simulate it in real time by the beginning of the next century.

Conclusion

I have suggested that future telecommunications systems have to implement information representations which are compatible with humans. I also emphasized that a holistic point of view is necessary to understand human functions. Such a view will provide the means to mediate, equalize and enhance human communications.

Acknowledgement

I appreciate discussions and proof reading for Malcolm Crawford, and other colleagues for comments and discussions.

References

[1] A. S. Bregman: "Auditory Scene Analysis," MIT Press, (1990).

[2] A. S. Bregman: "Auditory Scene Analysis: hearing in complex environments," in Thinking in Sounds, (Eds. S. McAdams and E. Bigand), pp.10-36, Oxford University Press, (1993).

[3] M. Kawato, H. Hayakawa and T. Inui: "A Forward-inverse Optics Model of Reciprocal Connections between Visual Cortical Areas," Network, 4, pp.415-422 (1993).

[4] D. Marr: "Vision," Freeman, (1982).

[5] D. T. Kemp" "Stimulated Acoustic Emissions from within the Human Auditory System," J.A.S.A., vol.64, pp.1386-1391 (1978).

[6] R. M. Warren: "Auditory Perception: A New Synthesis," Pergamon, (1982).

[7] W. B. Warr: "Organization of Olivocochlear Efferent Systems," in The Mammalian Auditory Pathway: Neuroanatomy, (D. Webster, A. Popper and R. Fay eds.), Springer-Verlag (1992).

[8] J. F. Brugg: "An Overview of Central Auditory Processing," in The Mammalian Auditory Pathway: Neurophysiology, (A. Popper and R. Fay eds.), Springer-Verlag (1992).

[9] H. McGurk and J. McDonald: "Hearing Lips and Seeing Voices," Nature, 264, pp.746-748 (1976).

[10] K. Sekiyama and Y. Tohkura: "McGurk Effect in Non-English Listeners," J.A.S.A., vol.90, pp.1797-1805 (1991).

[11] B. C. J. Moore: "An Introduction to the Psychology of Hearing," Academic Press (1989).

[12] P. K. Kuhl, M. Tsuzaki, Y. Tohkura and A. Meltzoff: "Human Processing of Auditory-Visual Information in Speech Perception," ICSLP'94, S11.4, Yokohama (1994).

[13] R. S. Heffner and H. E. Heffner: "Sound Localization in Mammals," in D. Webster, R. Fay and A. Popper eds., The Evolutionary Biology of Hearing, pp.691-715, Springer-Verlag (1992).

[14] A. N. Meltzoff, P. K. Kuhl and M. K. Moore: "Perception, Representation, and the Control of Action in Newborns and Young Infants: Toward a new synthesis," in Newborn attention (M. Weiss and P. Zelazo, Eds.), pp. 377-411. Ablex (1991).

[15] A. Fernald and P. K. Kuhl: "Acoustic Determinants of Infant Preference for Motherese Speech," Infant-Behavior-and-Development, vol.10, pp.279-293 (1987).

[16] P. K. Kuhl: "Human Adults and Human Infnats Show a Perceptual Magnet Effects for the Prototypes of Speech Categories, Monkey Do Not," Percept. Psychophys., vol.50, pp.93-107 (1991).

[17] D. Werker and R. C. Tees: "Cross-Language Speech Perception: Evidence for Perceptual Reorganization During the First Year of Life," Infant Behavior and Development, vol.7, pp.49-63 (1984).

[18] R. Yamada and Y. Tohkura: "Perception of American English /r/ and /l/ by Native Speakers of Japanese," in Speech Perception, Production and Linguistic Structure, Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka, pp.155-174, IOS Press (1992).

[19] R. Yamada: "The Effects of Experimental Variables in the Perception of American English /r,l/ by Japanese Listners," Percept. Psychophys., vol.52, pp.376-392 (1992).

[20] R. Yamada: "Effect of Extended Training on /r/ and /l/ identification by Native Speakers of Japanese," J. Acoust. Soc. Am., 93, 4, Pt.2, p.2391 (1993).

[21] B. S. Lee: "Effects of Delayed Speech Feedback," J.A.S.A., vol.22, pp.824-826 (1950).

[22] D. Pisoni et.al.: "Some Acoustic Phonetic Correlates of Speech Production in Noise," Proc. IEEE ICASSP, pp.1581-1584 (1985).

[23] H. Kawahara: "Effects of Natural Auditory Feedback on Fundamental Frequency Control," ICSLP'94, S24-2, pp.1399-1402 (1994).

[24] J. Hennessy and N. P. Jouppi: "Computer Technology and Architecture: An Evolving Interaction," Computer, vol.24, pp.18-29 (1991).

[25] H. Moravec: "MIND CHILDREN," Harvard Univ. Press (1988).

[26] C. Koch and I Segev: "Methods in Neuronal Modeling," MIT Press, (1989).

Learning to recognise speech in noisy environments

Martin Cooke, Malcolm Crawford and Phil Green

Speech & Hearing Research
Department of Computer Science, University of Sheffield, UK

1 Speech perception and auditory scene analysis

In many accounts of speech perception, there is an assumption — usually implicit — that the "speech material" reaches the relevant processing centres of the brain virtually unscathed. Understandably, this presumption afflicts most engineering approaches to automatic speech recognition, where the input is assumed to consist of reasonable quality speech from a single speaker, unadulterated by other acoustic sources. Typical acoustic environments are decidedly less accommodating, yet listeners are able to communicate in all but the most adverse of conditions. Thus, in the everyday perception of speech, the auditory system must first solve the problem of sorting out an arbitrarily-cluttered acoustic environment. Bregman's *Auditory Scene Analysis* (1990) represents a comprehensive experimental/theoretical account of this remarkable aspect of auditory function. Bregman's claim is that auditory primitives — low-level representations of acoustic components — are grouped into perceptual representations which he calls *auditory streams*, each of which represents a coherent centre of description for an acoustic source. Components are assigned to a stream only if they obey some rule of auditory organisation to ensure their compatibility with other parts of the stream. Grouping rules investigated to date include onset/offset synchrony, common amplitude or frequency modulation, harmonicity, in addition to sequential laws such as continuity of timbre or spatial location.

What are the implications of this view? To answer this question requires an account of typical auditory stream composition. However, much of the experimental (and, of late, computational) effort has focused on understanding the conditions under which the laws of organisation act (e.g. the degree of onset asynchrony required for a component to be perceptually segregated). Less attention has been devoted to the putative characteristics and processing of auditory streams themselves. Computational modelling appears to offer more fertile ground for investigating this area. By implementing algorithms based on experimental accounts of grouping by harmonicity, common AM, onset/offset synchrony, timbre and the like, it is possible to study modelled auditory streams. A typical stream resulting from our own models (Cooke, 1993; Cooke & Brown, in press; Brown & Cooke, in press) is depicted in figure 1. What is immediately clear from this figure is the incomplete nature of the recovered speech stream[1].

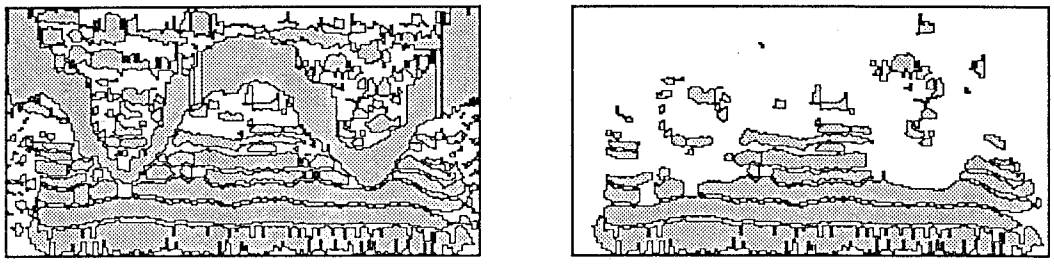


FIGURE 1. *Left*: Auditory time-frequency representation of an utterance mixed with a siren. *Right*: speech 'stream' produced by a model of ASA using the principle of grouping by pitch contour similarity.

If we accept the idea that all incoming signals are subject to an unconditional stage of auditory scene analysis, then we must entertain the hypothesis that *the representations employed in the normal processes of speech perception are conditioned on this potentially incomplete perceptual organisation.* In

1. Our model recovered about 40% of the energy associated with the speech source in this mixture. This is due in part to the paucity of grouping principles in the model. However, we hypothesise that there is an upper limit on the proportion of a single acoustic source that can be grouped purely by primitive processes. This limit will vary dependent upon the degree of acoustic clutter in the listener's environment. This is not a failure of mammalian auditory development. In other sensory domains, such as vision, the issue is clear cut: when an object is partially occluded by an opaque object, no data-driven process can guarantee successful completion of the pattern. Whilst acoustic sources are additive (transparent rather than opaque), there is both physiological (e.g. Sinex & Geisler, 1983) and psychoacoustic (e.g. Carlyon, this meeting) evidence that individual channels are dominated by components from single sources, rendering the analogy with visual occlusion more apt.

other words, speech perception is secondary to, and constrained by, a primary process which partitions the evidence into coherent but possibly fragmentary descriptions of acoustic sources.

This paper gives a brief overview of recent and ongoing work on computational studies which demonstrate the possibility of handling such fragmentary descriptions. Section 2 addresses the question of learning auditory-phonetic mappings in the presence of occluding 'noise'. Recognition of incomplete patterns using such maps is described in section 3, which also summarises results using a modification of the powerful hidden Markov model approach to automatic speech recognition. Section 4 extends the realism of these investigations through the use of incomplete patterns in which only spectral peaks remain, and section 5 shows how higher-level constraints from perceived auditory continuity can be incorporated into the processing framework.

2 Learning from incomplete auditory patterns

Auditory development proceeds in arbitrary acoustic environments — infants' post-natal acoustic experience is not confined to the anechoic chamber! It is of great interest to discover whether the possibility of incomplete source descriptions is an impediment to the learning process.

A simple simulation of the development of an auditory-phonetic mapping can be obtained via self-organising neural networks (Kohonen, 1984), in which input space similarities are mapped into topological similarity in a neural output grid. Such networks are trained in an unsupervised manner, and the categories thus obtained are classified *post hoc* using labelled speech material. Samad & Harp (1992) recently demonstrated a technique for adapting such networks to handle incomplete pattern vectors, and we have applied their method to the speech domain. In a series of experiments (Cooke, Green & Crawford, 1994), a self-organising network was trained using a 64 component auditory firing-rate representation. Data was derived from a portion of the TIMIT acoustic-phonetic database (Garofolo & Pallett, 1989). To simulate auditory scene analysis, portions of the data vector were deleted at random[2], with a probability which varied in 10 different conditions from 0.0 (no deletion) to 0.9 (90% deletion). The trained nets were then calibrated (i.e. one of 39 phone labels was attached to each output node) using the training set, and performance measured using a different set of labelled vectors. Figure 2 (left panel) shows label identification accuracy as a function of data deletion. The striking aspect of this graph is its flatness. The network trained on data with 90% of its components removed performs little worse than the network with all components present. This suggests that incomplete descriptions do *not* present serious difficulties to the learning process.

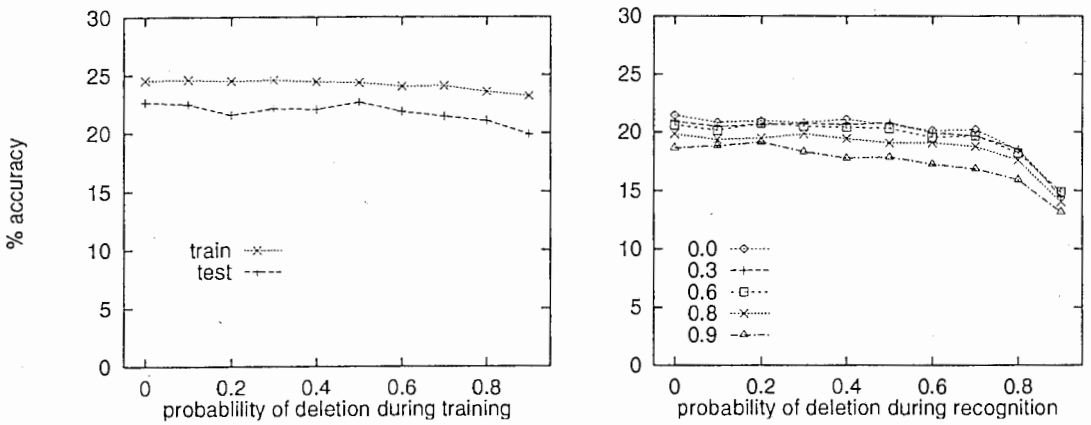


FIGURE 2. *Left*: recognition performance of Kohonen nets **trained** on vectors with various amounts of component deletion. *Right*: recognition performance for some of these nets using incomplete patterns **during recognition.**

3 Recognising incomplete patterns

Using Kohonen nets *trained* with partial data, we conducted a second series of experiments to determine the effect of missing data during *recognition*. Some of these results are presented in the right panel of figure 2. Encouragingly, recognition performance falls off gently with increasing component deletion, even for networks which have themselves been trained on partial data.

2. Of course, patterns which survive auditory scene analysis are likely to be anything but random, containing sequentially-correlated information (correlated deletions where the signal was dominated by a correlated masking source, and correlated data where the signal dominated its background). The use of non-random deletions is reported in section 4.

In a separate set of studies (Cooke, Green, Anderson & Abberley, 1994) we adapted continuous density hidden Markov models to handle incomplete observation vectors, and achieved a similar pattern of results. The left panel of figure 3 displays correctness ('hit rate') and accuracy ('hit rate' minus

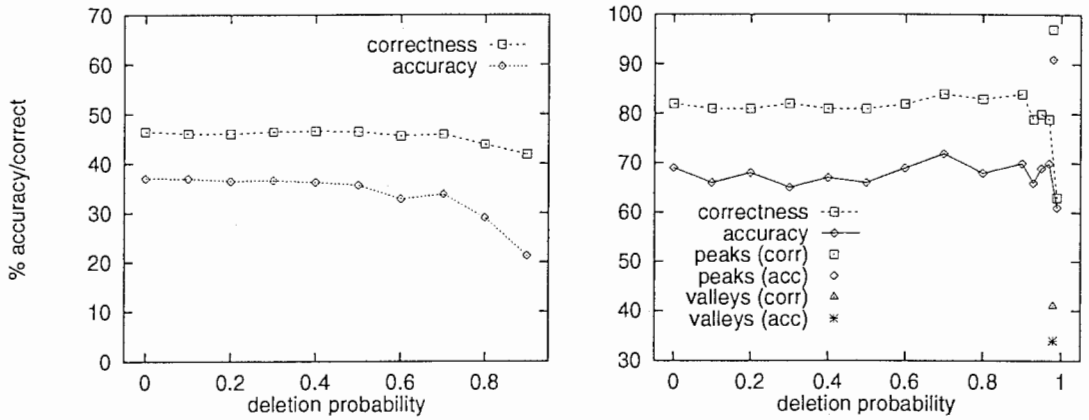


FIGURE 3. Recognition studies using HMMs. *Left*: Performance as a function of component deletion on a phone recognition task, using a auditory filterbank input representation. *Right*: Performance on a digit recognition task, comparing random component deletions (solid lines) and selective deletion, preserving peaks or valleys.

insertions and deletions) on a phone recognition task using the TIMIT database. The baseline level of performance is much higher, and could be further improved by more sophisticated models; however, it is not the level of performance which is of interest here, but the manner in which performance changes with increasingly incomplete input patterns.

4 Correlated deletions

As noted earlier, random component deletion is an inappropriate simulation of auditory scene analysis. In a more realistic approach, we compared the effect of random deletion with that of selective retention of all peaks or all valleys in the rate spectrum. The results, shown in the right panel of figure 3, favoured peaks over valleys, and, more interestingly, demonstrated better performance for the peak pattern than for the pattern with no deletions. Channels with peaks accounted for around 10-15% of the available data, yet better performance was obtained than when 100% of the data was presented. This result is explicable on the basis that any recognition approach which relies on the fine detail of the spectrum will be at the mercy of noise, particularly in the valleys. However, the result is principally of interest for another reason: a peak-based pattern would be automatically derived by auditory scene analysis (at least for peaks corresponding to resolved harmonics in the lower frequencies), and does not require the existence of an explicit peak picker.

These studies suggest a way out of a difficulty identified by many researchers who have attempted to use an auditory front-end to an automatic speech recogniser, viz: sensitivity to F0 (e.g. Beet, 1990).The problem is this: resolution of harmonics in the F1 (and often F2) region, apparent in most auditory representations, leads to an excitation pattern which is highly variable across tokens of the same voiced sounds, due to F0 differences across the tokens. A recognition technique such as that proposed here solves the problem since the harmonic pattern obtained by auditory scene analysis determines those frequencies at which the spectrum should be sampled, rather than defining a pattern to be matched *per se*. This is not a novel idea: Assman & Summerfield (1989) used a similar technique to identify formant frequencies in excitation patterns which contained resolved harmonics. However, their 'peak' metric required explicit identification of each formant, whilst the partial matching approach outlined here operates without such a requirement.

5 Higher-level constraints

Whilst grouped material might form a primary 'key' into speech schemas, there is reason to assume that the recognition process itself has access to something more like a complete auditory scene. Studies of perceived auditory continuity (Warren, 1970, Bashford & Warren, 1987, Warren *et al.*,1994) suggest that the full spectral profile places constraints on valid schema-matches. Specifically, 'auditory induction' operates only if sufficient energy exists in the occluded regions to account for the induced pattern.

This property can be used to refine the process of partial matching through a modification to the Kohonen net recognition algorithm in which a penalty is applied for any missing components whose maximum value (provided by the level in the mixture) falls below that expected on the basis of the components which are present. In other words, the incomplete vector defines a possible matching

pattern, whose values (weights in the Kohonen net) represent a prediction of the expected energy at that spectral place. If there is insufficient energy in the mixture at that place, then there is certainly insufficient in any source which makes up the mixture. Figure 4 presents the results of adding this constraint to the

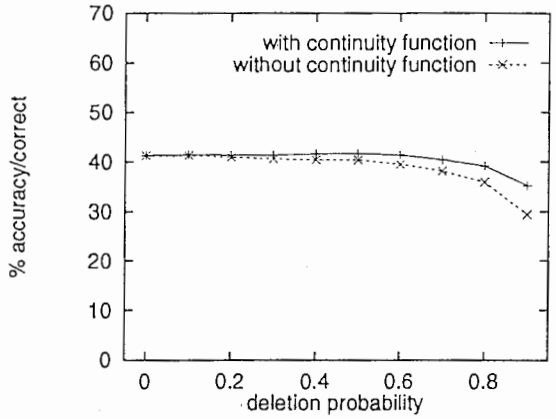


FIGURE 4. Recognition accuracy vs. probability of deletion for 2 recognition algorithms. The "without continuity function" curve employs the standard distance measure whilst "with continuity function" includes modifications suggested by the continuity effect as described in the text.

recognition algorithm, and clearly shows a further flattening of the recognition curve as the probability of deletion increases. More sophisticated exploitation of illusory continuity is possible too: for instance, in Cooke & Brown (1993), the value of missing harmonic components was predicted in a purely bottom-up fashion. Likewise, syllabic, lexical and even phrasal knowledge could be utilised to better estimate the level of missing components prior to comparison with the overall level of the mixture.

6 Discussion

This paper reports on computational studies which demonstrate that recognition performance of self-organising neural networks and hidden Markov models is barely affected by random speech component deletion, and can be enhanced by selective deletion which retains spectral peaks. Furthermore, development of the representational substrate for recognition can still proceed within this regime. We are currently completing the processing pathway by using an auditory scene analysis front-end rather than simulated deletions. We intend to develop further the sophisticated hidden Markov model approach to *learn* from incomplete data.

The potential of these investigations for speech technology in adverse conditions is clear; what is more speculative is their message for speech perception. At the very least, they provide a constructive demonstration that less-than-perfect acoustic source characterisation is no real impediment to learning and recognition. This goes some way to providing a coherent framework for a variety of experimental studies which have provided evidence for what could be called the 'minimalist' school of speech perception e.g. the relative importance of formant frequencies in vowel perception (Carlson, Granstrom & Klatt, 1979); the least spectral contrast required to detect formants (Lea, 1992); vowel identification from single formants (Sawusch, 1991). Additionally, the current studies explain how the relevant vowel schema themselves might be learnt, an aspect which was not addressed in the investigations reported above.

A bolder speculation is that the patterning of sounds is conditional upon prior scene analysis (e.g. use of VOT differences in stop consonant identification might have a basis in grouping by onset synchrony). We have already mentioned the distinction between data-driven primitive grouping and schema-based processing, and suggested that the principle of perceived auditory continuity can be exploited to allow these two processes to interact. One model which extends this view is the notion that *unconditional primitive grouping provides a primary route into schema memory, and the most active schemas thus invoked verify themselves against the auditory scene as a whole* (using mechanisms such as that described in section 5). Speech perception in relatively quiet environments would be dominated by accurate schema-invocation from primitive grouping. In adverse conditions, a larger number of schemas would compete to explain the evidence, necessitating more processing. An extreme example of this is the sine-wave speech stimuli employed by Remez *et al.* (1981) in which all cues for primitive grouping are systematically eliminated. In the model proposed here, each separate formant would form a primary key into speech (and nonspeech) schemas, and, although each formant track would not usually uniquely identify a syllable sequence, it would invoke certain schemas more than others. Top-down processing of these schemas would lead to a reasonable chance of making sense of such stimuli. Of course, an additional cue for grouping, such as comodulation of the formants, would drastically limit the search, and

this is one interpretation of the dramatic improvement in listeners' ability to recognise such sentences found by Carrell & Opie (1992).

In conclusion, these investigations suggest that the treatment of speech perception as a secondary pattern recognition process occurring subsequent to an initial organisation of sound components is not only feasible, but may lead to representations for both learning and recognition which are different from those traditionally considered important in the field of speech perception.

References

P.F. Assman & Q. Summerfield (1989), "Modeling the perception of concurrent vowels: vowels with the same fundamental frequency", *JASA*, **85** (1), 327-338.

J.A. Bashford Jr. & R. M. Warren (1987), "Multiple phonemic restorations follow the rules for auditory induction", *Perception & Psychophysics*, **42** (2), 114-121.

S.W. Beet (1990), "Automatic speech recognition using a reduced auditory representation and position-tolerant discrimination", *Computer Speech & Language*, **4** (1), 17-33.

A.S. Bregman (1990), *Auditory Scene Analysis*, MIT Press.

G.J. Brown & M.P. Cooke (in press), "Computational auditory scene analysis", *Computer Speech & Language*.

R. Carlson, B. Granstrom & D. Klatt (1979), "Vowel perception: the relative perceptual salience of selected acoustic manipulations", STL-QPSR 3-4/79, RIT Stockholm.

T.D. Carrell & J.M. Opie (1992), "The effect of amplitude comodulation on auditory object formation in sentence perception", *Perception & Psychophysics*, **52**, 437-445.

M.P. Cooke, P.D. Green and M.D. Crawford (1994), "Handling missing data in speech recognition", *International Conference on Speech and Language Processing*, Yokohama.

M.P. Cooke and G.J. Brown (in press), "Separating simultaneous sound sources: issues, challenges and models", *Speech Recognition and Speech Synthesis* (ed. E. Keller), John Wiley & Sons.

M.P. Cooke and G.J. Brown (1993), "Computational auditory scene analysis: Exploiting principles of perceived continuity", *Speech Communication*, **13**, 391-399.

M.P. Cooke, P.D. Green, C. Anderson & D. Abberley (1994), "Recognition of occluded speech by hidden Markov models", University of Sheffield Department of Computer Science Technical Report TR-94-05-01 (submitted to *Computer Speech & Language*).

M.P. Cooke (1993), *Modelling Auditory Processing and Organisation*, Cambridge University Press.

J.S. Garofolo & D.S. Pallett (1989), "Use of the CD-ROM for speech database storage and exchange", *Proc. Euro. Conf. Speech Communication and Technology*, Paris, 309-315.

T. E. Kohonen (1984), *Self-Organisation and Associative Memory*, Springer.

T. E. Kohonen, J. Kangas & J. Laaksonen (1992), "SOM_PAK, The Self-Organizing Map Program Package, Version 1.2", Laboratory of Computer and Information Science, Helsinki University of Technology, Finland.

A.P. Lea (1992), "Auditory modelling of vowel perception", Ph.D. Thesis, University of Nottingham.

R.E. Remez, P.E. Rubin, D.B. Pisoni & T.D. Carrell (1981), "Speech perception without traditional speech cues", *Science*, **212**, 947-950.

T. Samad & S.A. Harp (1992), "Self-organisation with partial data", *Network*, **3**, 205-212.

D.G. Sinex & C.D. Geisler (1983), "Responses of auditory-nerve fibers to consonant-vowel syllables", *JASA*, **73** (2), 602-615.

R.M. Warren (1970), "Perceptual restoration of missing speech sounds", *Science*, **167**, 392-393.

R.M. Warren, J.A. Bashford Jr., E.W. Healy & B.S. Brubaker (1994), "Auditory induction: Reciprocal changes in alternating sounds", *Perception & Psychophysics*, **55** (3), 313-322.

Acknowledgements

This work was supported by SERC Image Interpretation Initiative Research Grant GR/H53174, a study visit grant from ATR to Malcolm Crawford and travel grants from the Royal Society and the Royal Academy of Engineering to Phil Green and Martin Cooke. Guy Brown kindly made his models available. Kohonen net simulation employed the public domain SOM_PAK software (Kohonen, Kangas & Laaksonen, 1992).

A Neurobiological Perspective on Speech Production

Vincent L. Gracco
Haskins Laboratories

Introduction

Speech is one of man's most distinguishing traits. At the core of this evolutionary advance is the human nervous system represented as an processing machine constantly receiving, integrating, and exchanging information through a variety of sensorimotor channels. A premise of this short report is that in order to understand speech communication it is essential to understand the basic neural components and the sensorimotor processes operating on them. The focus, then, is to outline a conceptual framework for speech production that is rooted in neurobiology. An assumption of this approach is that speech should not be considered an isolated property of the nervous system but an integrative and interactive process that uses all aspects of neural functioning to accomplish information exchange. Moreover, speaking as a motor act, should not be viewed as a separate "low level" behavior isolated from "high level" cognitive functioning. Rather, speech production can be viewed as a model of nervous system functioning and principles associated with speech should be reflective of and ultimately reduce to general nervous system functions for most kinds of human behavior.

Neural Substrate and Some Functional Considerations

The neural mechanisms for speech can be viewed from two perspectives; the neuroanatomical substrate and the nature of the control mechanisms. The neural substrate for speech has been identified from a variety of sources including human mapping studies using electrical stimulation (Penfield & Roberts, 1959; Ojemann, 1983; Mateer, 1983) and neuroanatomical studies of nonhuman primates (Muakassa & Strick, 1979; Woolsey, Settlage, Meyer, Sencer, Pinto Hamuy, & Travis, 1952). A number of cortical and subcortical regions have been identified in which a representations of the vocal tract are found (see Gracco & Abbs, 1987; Barlow & Farley, 1989 for reviews). Cortical regions with vocal tract representations include the primary motor and sensory areas (MI and SI, respectively), the so-called nonprimary motor areas including supplementary motor area (SMA) and premotor area (PM; lateral precentral cortex), and a posterior parietal region. The general PM area and posterior parietal regions (including portions of the temporal region in man) comprise the areas associated with Broca's and Wernicke's areas respectively (Penfield & Roberts, 1959). Extensive subcortical representations can also be found in the cerebellar cortex, deep cerebellar nuclei and regions of the basal ganglia. An interesting aspect of these representations is their overall connectivity. For example, different cortical areas are connected to different subcortical structures and contain projections from or to distinct and (relatively) non-overlapping regions of the thalamus as well as subcortical structures (basal ganglia and cerebellum). The PM area receives input from the deep cerebellar nuclei via the thalamus and projects to the primary motor area (MI) as well as contributing direct descending projections to brain stem nuclei. Similar segregated extrinsic connections are found for regions of the basal ganglia and SMA. In addition, there are rather dense projections from parietal areas to the motor and premotor cortical areas as well as temporal regions and descending projections to brain stem nuclei (Gracco & Abbs, 1987). A summary of neuroanatomical data reported by Schell and Strick (1984) suggests that large regions of the cortex and subcortex are interconnected and

maintain relatively segregated networks that ultimately converge at the output. These diverse neural areas, which represent large regions of the nervous system, display an extrinsic organization consistent with the concept of neural modules hypothesized by Mouncastle suggesting distributed processing functions (Mountcastle, 1978). It should be noted that these large scale networks all have access to peripheral sensory information from somatic receptors as well as the visual and auditory receptors and therefore display "reentrant" characteristics such that changes in one system allows changes or readjustment in all convergent systems (Edelman, 1987). Generally speaking the neuroanatomy underlying speech and language is quite complex and highly specialized to receive, integrate and act on the external and internal environment of the organism.

A primary source of insight on the nervous system organization for speech and language comes from neurological disorders. From a synthesis of various observations some general conclusions can be drawn. A surprising characteristic of almost all lesions involving the central nervous system is the associated motor impairments accompanying cognitive or linguistic impairments. It appears as suggested by Jackson (1875) that the so called higher centers of the nervous system may be extensions of the lower nervous centers which represent impressions and movements. Consistent with the neuroanatomical substrate outlined above damage to the cerebellum and/or PM area often result in sensorimotor impairments of some similarity at least to acoustic and perceptual examination (Kent & Rosenbeck, 1982). Damage to either of these regions often produces a breakdown in speech that can be characterized by a disruption of the smooth timing of sequential speech movements. Cerebellar patients often show a decomposition of movement as though the various parts of a complex movement had to be thought out one by one. Dysmetria is also a characteristic of cerebellar damage suggesting that the ability to integrate somatic and visual information to produce appropriately calibrated actions has been affected. These symptoms are generally consistent with those associated with Broca's aphasia (due to anterior premotor lesions). For example, electrical stimulation of the PM area, which receives output from the deep cerebellar nuclei via the thalamus causes speech arrest (Penfield & Roberts, 1959) and an inability to sequence multiple speech movements (Mateer, 1983). Because of the limited data available it is not clear just how similar damage to these regions is but it can be suggested that there may be considerable overlap. A similar suggestion can be made regarding the impairments associated with basal ganglia and SMA damage. Basal ganglia damage, characterized by Parkinson's disease, often results in speech characterized by imprecise consonant production, mono-pitch and loudness, and articulator movements that are reduced in amplitude and slow. SMA damage results in speech impairments ranging from one extreme, total speech arrest, to imprecise articulation. Finally, damage to the posterior portion of the brain (Wernicke's area) produces speech and language impairments consistent with a role in the complex processing of multimodal input and the contribution of that processing to the final motor output. Patients with damage to posterior portions of the brain display output impairments that have been suggested to reflect inappropriate phonological selection compared to the more phonetic errors exhibited by anterior (Broca's) aphasics (Blumstein, 1981). Thus there appears to be some general functions associated with the hypothetical distributed processing modules known to represent vocal tract sensorimotor structures.

Speech Production Units--Fundamental Representations

An issue of paramount importance in speech production and motor behavior in general is to determine the level at which neural control is being exerted and hence the elementary units of the targeted behavior. Classical analyses of the integration of neuromuscular behavior (e.g.., Weiss, 1941) strongly suggests that a general principle of neural organization is that neither the physical variables nor the particular subunits of a motor ensemble (individual muscles) are the key elements in the central neural constructs for motion. As suggested by

Bernstein almost 30 years ago (Bernstein, 1967) functional behavior involving multiple degrees of freedom is mastered by organizing the process as a whole. That is, the control of multiarticulate behaviors involving high dimensionality is simplified by constraining actions of functionally related effectors into smaller controllable units (see also Fowler, Rubin, Remez, & Turvey, 1980; Gracco, 1988; Kelso, 1986; Turvey, 1977; Saltzman & Kelso, 1987). However, while it is clear that speech is organized at a task level, an unresolved issue for production models is the level of that organization. Two different kinds of empirical observations have been made that reflect directly on the level of control exerted by the nervous system during motor speech. One such observation is the response of speech articulators to a mechanical perturbation. If a bite block is placed between the teeth restricting the jaw's position during vowel production, speakers make adjustments in the tongue position and shape to produce a vowel with perceptually similar characteristics. Perturbations to the motion of a moving speech articulator display related characteristics with responses demonstrating task-specific patterns of compensation. Mechanical perturbation to the lips result in compensatory changes in the lips and jaw (Abbs & Gracco, 1984; Gracco & Abbs, 1985; 1988) and the larynx (Löfqvist & Gracco, 1991; Munhall, Löfqvist, & Kelso, 1994); jaw loads result in compensatory changes in the tongue (Kelso et al., 1984), lips (Folkins & Abbs, 1975; Shaiman, 1989), and velum (Kollia et al., 1992). Task specific responses are observed when an articulator is actively involved in the sound segment being produced but not when an articulator is not involved (Kelso et al., 1984; Shaiman, 1989). These studies suggest that the task-specific requirement for observing compensation to perturbation is the physiological composition of the phonetic segment being produced. Observations of the relative timing of articulators is also consistent with the general observation that articulator actions are organized in functional aggregates. Upper lip, lower lip, jaw, velum and laryngeal motion has been shown to display systematic covariation in timing (Gracco, 1988; Gracco, 1994; Gracco & Abbs, 1986; Gracco & Löfqvist, 1994; Kollia, Gracco, & Harris, submitted) with the timing among articulators stronger within a phonetic segment than across phonetic segments. These observations have been taken to offer support for phonetic units as the level of neural representation for speech reflecting categorically invariant neuromotor patterns (Gracco, 1991; Gracco & Löfqvist, 1994).

Phoneme based models of speech perception and production have a long history. One problem that has plagued theoretical perspectives emphasizing phonetic representations for speech is that the acoustic correlates of a given phoneme and by inference vocal tract configurations exhibit variability from a number of sources including context, speaking rate, stress, etc. Such observations indicate an important property of the speech production system regarding the control of speech movements and the degree of control precision for speech. The spatial variability that characterizes speech motion can be interpreted as reflecting loosely specified goals in an abstract task space (Abbs, Gracco, & Cole, 1984; Saltzman & Munhall, 1989). That is, the details of a task are not specified except in a general sense, with mechanisms available to assure accurate perception. As pointed out by von Nuemann (1958) the nervous system is an analog device that is ideally suited for reliable operation not precision. In this context it can be suggested that articulatory performance is good enough without incurring excessive costs. An example of the degree to which speech movements need only be loosely controlled can be found in recent simulation and synthesis results reported by Gay, Boe, & Perrier (1992). Parametric manipulation of vocal tract cross sectional area and constriction location was used to determine the acoustic and perceptual boundaries of certain isolated vowels. It was shown that the formants for each of the vowels were most sensitive to changes in cross sectional area compared to constriction location. Vowel perception, however, was insensitive to both manipulations. The results from Gay et al. (1992) were somewhat at odds with the notion of the quantal characteristics of speech (Stevens, 1972; 1989) suggesting rather that quantal regions for vowels may not necessarily be avoided because of the tolerance of the

perceptual system. From these results it was concluded that the speech production mechanism has "considerable latitude" in specifying the articulatory targets (Gay et al., 1992). An additional interpretation is that the perceptual system is sufficiently tolerant to accept a wide range of variation effectively relaxing the control required for speech production.

Sequencing of Speech Motor Actions

Speech is more than the specification of characteristic motor patterns or phonetic segments adjusted for context. An important consideration in speech production is the sequencing of vocal tract actions into communicatively meaningful units of production. While speech is a specialized human function, the view taken here is that it is one of many important brain functions and any theoretical account must adhere to principles that are shared by other similar behaviors. If one accepts the premise that the human brain has evolved from earlier brains, (based on the need to predict and control species-specific events in the environment), then supposing that more complex, higher-level mechanisms, developed from less complex, lower level mechanisms, is a logical extension. In this regard, the fundamental sequential nature of speech may be similar to other fundamentally sequential behaviors such as locomotion respiration, or mastication (Lashley, 1951; Kozhevnikov & Chistovich, 1965). This is not to suggest that speech shares specific motor patterns with other rhythmic behaviors. Rather, they may share similar principles for their implementation as well as adhere to similar organizational principles although they will be adapted to specific task requirements (e.g., communication) and effector properties (see Grillner, 1982; Gracco, 1990; Kelso & Tuller, 1984). For more automatic behaviors such as mastication and locomotion, central rhythm generators have been identified which produce behavior-specific rhythmic motor output similar in form and function to those identified in lower vertebrates. Differences in muscle activity and movement patterns for speech, chewing, and respiration clearly indicate that the same central pattern generator does not underlie all behaviors. Rather, a number of observations are consistent with the presence of some kind of rhythm generating mechanism as the basis for sequential speech motor adjustments. For example, mechanical perturbation of speech movement sequences result in an increase or decrease in the movement cycle frequency, dependent on the phase of the movement during which the load is applied (Gracco & Abbs, 1988; 1989; Saltzman, Kay, Rubin, & Kinsella-Shaw, 1991; Löfqvist, Saltzman, Kinsella-Shaw, Rubin, & Kay, 1994; Saltzman, Löfqvist, Kinsella-Shaw, Rubin, & Kay, 1992). One interpretation of these results is that a rhythmic mechanism is the foundation for the serial timing of speech movements, and that the mechanisms is modifiable, not stereotypic (Gracco, 1990; 1991). Such a central rhythm generator would provide a framework for the sequencing of production units that, in turn, modulate the instantaneous frequency of the oscillator, based on intrinsic, phoneme-specific requirements (Gracco, 1990; 1991; 1994). An important consequence of incorporating a central rhythm generator into a speech production model is the ability to explain rate, stress, and final lengthening changes with manipulation of a single mechanism; global and local changes in rhythmic frequency. Moreover, the rhythmic nature of the output assists in one of the characteristic problems related to speech perception; segmentation (Cutler & Mehler, 1993; Lashley, 1951; Martin, 1972). The rhythmic modulation of sound production provides the perceptual system with a framework for sampling and parsing the input. The breakdown in the rhythmic structure of speech associated with a number of different speech motor disorders strongly suggests that the underlying rhythm is a network property rather than residing in a specific neuroanatomical location (Kent & Rosenbek, 1982).

Summary

The model outlined rests on a number of assumptions about nervous system function and associated organizational principles that while specific to speech production, may generalize to many other behaviors. First, the level of control for speech is minimally at a level that reflects the smallest functional unit of speech, the phonetic gesture. Similarly, these units are ultimately organized into larger units on the order of syllables or stress units which are no doubt organized into larger units such as lexical items. This suggests that speech production is a nested process with function, at each level, as the organizing principle. Speech movements are contextually variable suggesting that vocal tract configurations for the phonetic segments are only loosely specified and the degree of variation is dictated by the perceptual requirements of each individual phonetic segment or class of speech sounds. Additional mechanisms operate to sequence the units into aggregates that are also sufficiently flexible reflecting substantial degrees of variation. A fundamental principle that emerges is that communication is a synthetic and stochastic process organized at multiple levels in parallel. As a result no one movement component or signal attribute is solely responsible for information transfer; the neural control of speech relies on flexible processes and reliable performance rather than invariant principles and rigid tolerances. Speech production (and presumably speech perception) is a distributed and integrative process systematically operating on a number of time scales and along a number of sensorimotor dimensions transforming intent into a series of coarticulating sound sequences.

References

Abbs, J.H., & Gracco, V.L. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of the lip during speech. Journal of Neurophysiology, 51(4), 705-723.

Abbs, J.H., Gracco, V.L., & Cole, K.J. (1984). Control of multimovement coordination: Sensorimotor mechanisms in speech motor programming. Journal of Motor Behavior, 16, 195-232.

Barlow, S. M., & Farley, G. R. (1989). Neurophysiology of speech. In D.P. Kuehn, M. L. Lemme, & J. Baumgartner, (Eds.), Neural Bases of Speech, Hearing, and Language (pp. 146-200). Boston: College-Hill Press.

Bernstein, N. (1967). The co-ordination and regulation of movements. New York: Pergamon Press.

Blumstein, S. E. (1981). Neurolinguistic disorders: Language-brain relationships. In, S. B. Filskov and T. J. Boll (Eds.), Handbook of Clinical Neuropsychology. New York: Wiley.

Cutler, A., & Mehler, J. (1993). The periodicity bias. Journal of Phonetics, 21, 103-108.

Edelman, G. M. (1987). Neural Darwinism. New York: Basic Books, Inc.

Folkins, J. W., & Abbs, J. H. (1975). Lip and jaw motor control during speech: Responses to resistive loading of the jaw. Journal of Speech Hearing Research, 18, 207-220.

Fowler, C. A., Rubin, P., Remez, R. E. & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In B. Butterworth, (Ed.), Language production (pp. 373-420). New York: Academic Press.

Gay, T., Boe, L.-J., & Perrier, P. (1992). Acoustic and perceptual effects of changes in vocal tract constrictions for vowels. Journal of the Acoustical Society of America, 92, 1301-1309.

Gracco, V. L. (1994). Some organizational characteristics of speech movement control. Journal of Speech and Hearing Research, 37, 4-27.

Gracco, V.L. (1991). Sensorimotor mechanisms in speech motor control. In, H. Peters, W. Hultsijn, & C.W. Starkweather (eds.), Speech Motor Control and Stuttering, North Holland: Elsevier, 53-78.

Gracco, V.L. (1990) Characteristics of speech as a motor control system. In, G. Hammond (Ed.), Cerebral Control of Speech and Limb Movements, North Holland: Elsevier, 3-28.

Gracco, V.L. (1988) Timing factors in the coordination of speech movements. Journal of Neuroscience, 8(12), 4629-4639.

Gracco, V. L., & Löfqvist, A. (1994). Speech motor organization and control: Evidence from lip, jaw, and laryngeal interactions. Journal of Neuroscience.

Gracco, V.L., & Abbs, J.H. (1989). Sensorimotor characteristics of speech motor sequences. Experimental Brain Research, 75, 586-598.

Gracco, V.L., & Abbs, J.H. (1988). Central patterning of speech movements. Experimental Brain Research, 71, 515-526.

Gracco, V.L., & Abbs, J.H. (1987). Programming and execution processes of speech movement control: Potential neural correlates. In E. Keller & M. Gopnik (Eds.), Symposium on Motor and Sensory Language Processes. New Jersey: Lawrence Erlbaum Associates, Inc., 163-201.

Gracco, V.L., & Abbs, J.H. (1986). Variant and invariant characteristics of speech movements. Experimental Brain Research, 65(1), 156-166.

Gracco, V.L., & Abbs, J.H. (1985). Dynamic control of the perioral system during speech: Kinematic analyses of autogenic and nonautogenic sensorimotor processes. Journal of Neurophysiology, 54(2), 418-432.

Grillner, S. (1982). Possible analogies in the control of innate motor acts and the production of sound in speech. In S. Grillner, B. Lindblon, J. Lubker, & A Persson (Eds.), Speech motor control (pp. 217-230). Oxford: Pergammon Press.

Jackson, J. Hughlings (1875). Clinical and physiological researches on the nervous system. (Contains a reprinting of "On the anatomical and physiological localisation of movement in the brain," first published in Lancet, 1873, i:84-85). Churchill: London.

Kelso, J. A. S., & Tuller, B. (1984). Converging evidence in support of common dynamic principles for speech and movement coordination. American Journal of Physiology,15, R928-R935.

Kelso, J. A. S., Tuller, B, V.-Bateson, E., & Fowler, C. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. Journal of Experimental Psychology: Human Perception and Performance, 10, 812-832.

Kelso, J. A. S. (1986). Pattern formation in speech and limb movements involving many degrees of freedom. In H. Heuer & C. Fromm (Eds.), Generation and modulation of action patterns (pp. 105-128). Berlin: Springer-Verlag.

Kent, R. D., & Rosenbek, J. C. (1982). Prosodic disturbance and neurologic lesion. Brain and Language, 15, 259-291.

Kollia, H. B., Gracco, V. L., & Harris, K. S. (submitted). Lip, jaw, velar coordination during speech. Journal of the Acoustical Society of America

Kollia, Gracco & Harris (1992). Functional organization of velar movements following jaw perturbation. Journal of the Acoustical Society of America, 91(2), 2474

Kozhevnikov, V., & Chistovich, L. (1965). Speech: Articulation and perception. Joint Publications Research Service, 30,453; U.S. Department of Commerce.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A.. Jeffress (Ed.), Cerebral mechanisms in behavior: The Hixon symposium. New York: Wiley.

Lindblom, B., Lubker, J., Gay, T., Lyberg, P., Branderal, P., & Holgren, K. (1987). The concept of target and speech timing. In R. Channon & L. Shockery (Eds.), In honor of Ilse Lehiste (pp. 161-181). Dordrecht, The Netherlands: Foris Publications.

Löfqvist, A., Saltzman, E., Kinsella-Shaw, J., Rubin, P., Kay, B (1994). Phase resetting in speech. II. Discrete utterances. Journal of the Acoustical Society of America, 95, (5;Pt.2), 2823.

Löfqvist, A., & Gracco, V. L. (1991). Discrete and continuous modes in speech motor control. PERILUS, XIV, 27-34.

Motor Behavior, 16(2), 195-232.

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. Psychological Review, 79, 487-509.

Mateer, C. A. (1983). Motor and perceptual functions of the left hemisphere and their interactions. In S. J. Segalowitz, (Ed.). Language functions and brain organization. New York: Academic Press, pp. 145-170.

Mountcastle, V. B. (1978). An organizing principle for cerebral function: The unit module and the distributed system. In, G. M. Edelman & V. B. Mountcastle, (Eds.), The mindful brain: Cortical organization and the group-selective theory of higher brain function. Cambridge: MIT Press, pp. 7-50.

Muakassa, K. F., & Strick, P. L. (1979). Frontal lobe inputs to primate motor cortex: evidence for four somatotopically organized "premotor" areas. Brain Research, 177, 176-182.

Munhall, K., Löfqvist, A., & Kelso, J. A. S. (1994). Lip-larynx coordination in speech: effects of mechanical perturbations to the lower lip. Journal of the Acoustical Society of America, 96, 3605-3616.

Ojeman, G. A. (1983). Brain organization for language from the perspective of electrical stimulation mapping. The Behavioral and Brain Sciences, 6, 189-230.

Penfield, W., & Roberts, L. (1959). Speech and brain mechanisms. Princeton, N. J.: Princeton Univ. Press.
Saltzman, E. L. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In, H. Heuer & C. Fromm (Eds.), Generation and modulation of action patterns (pp. 129-144). Berlin: Springer-Verlag, .
Saltzman, E. L., & Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. Psychological Review, 94, 84-106.
Saltzman, E. L., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. Ecological Psychology, 1, 333-382.
Saltzman, E., Löfqvist, A., Kinsella-Shaw, J., Rubin, P, & Kay, B (1992). A perturbation study of lip-larynx coordination. In Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP '92): Addendum, Edmonton, Alberta, Canada: The University of Alberta.
Saltzman, E., Löfqvist, A., Kinsella-Shaw, J., Rubin, P, & Kay, B (1994). Phase resetting in speech. I. Repetitive utterances.Journal of the Acoustical Society of America, 95, (5;Pt.2), 2823.
Schell, G. R. , & Strick, P. L. (1984). The origin of thalamic inputs to the arcuate premotor and supplementary motor areas. Journal of Neuroscience, 4, 539-560.
Shaiman, S. (1989). Kinematic and electromyographic responses to perturbation of the jaw. Journal of the Acoustical Society of America, 86, 78-87.
Stevens, K. N. (1972). On the quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David & P. B. Denes (Eds.), Human communication: A unified view. New York: McGraw-Hill, pp. 51-66.
Stevens, K. N. (1989). On the quantal nature of speech. Journal of Phonetics, 17, 3-45.
Turvey, M. T. 1977. Preliminaries to a theory of action with reference to vision. In, R. Shaw & J. Bransford (Eds.), Perceiving, Acting and Knowing: Towards an Ecological Psychology. Hillsdale: Lawrence Erlbaum.
von Nuemann, J. (1958). The computer and the brain. New Haven: Yale University Press.
Weiss, P. (1941). Self-differentiation of the basic patterns of coordination. Comparative Psychology Monograph, 174, 1-96.
Woolsey, C. N., Settlage, P. H., Meyer, D. R., Sencer, W., Pinto Hamuy, T., & Travis, A. M. (1952). Patterns of localization in precentral and "supplementary" motor areas and their relation to the concept of a premotor area. Association for Research in Nervous and Mental Disease, 30, 238-264.

Somatoneural Relation in the Auditory-Articulatory Linkage

Kiyoshi Honda

ATR Human Information Processing Research Laboratories
(2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan)

1. Introduction

The aim of this paper is to propose a functional neuroanatomical linkage between speech perception and production which serves as a basic mechanism to facilitate human speech communication. Exploring this proposition through experimental procedures is practically difficult because the linkage is concealed within the human brain. A possible way to reveal this human-specific function is to consider the shapes of the speech organs and their neural representations. The comparative morphology of the relevant peripheral systems suggests that the establishment of human speech communication is primarily due to the elaboration of speech production capabilities. The development of the auditory mechanism, in contrast, appears to show a continuation from the original form, permitting the common function of sound perception and localization. These accounts indicate that the reorganization of the neural maps of the body associated with the evolution of the speech organs may be the factor that binds speech production and perception together. In this paper, morphological and physiological studies of the mechanisms of vowel articulation and fundamental frequency (F0) control are reviewed with reference to the body-brain relationship. The results of the studies lead us to hypothesize the speech module in the brain consisting of sensorimotor representations of vowel and F0. The articulatory perspective on the speech module further emphasizes vowel-F0 integration in the auditory-articulatory linkage.

2. Somatoneural Relation and Human Specific Form of Speech Organs.

Neural connectivity is dependent on the changing size and form of an animal's body (Purves, 1988). The concept of this "somatoneural relation" is generally supported by biologists' observation on the form of vertebrates' body and its representation in the brain. In the animals which demonstrate a drastic metamorphosis, e.g., amphibians, the musculoskeletal system shows a significant difference between larval and adult forms. However, the neural connections between the brain and the peripheral organs remain unchanged during the metamorphosis. This fact suggests that the neural function must alter according to changes in the body shape. The human speech production system demonstrates a unique form in comparison with other primates, as shown in Fig. 1. The specific form of the human speech organs includes a short oronasal prominence, a round tongue, a wide pharynx, and a low position of the larynx. In particular, the separation of the larynx from the tongue in humans provides a morphological advantage to allow continuous vocal fold vibration during large displacements of the tongue and jaw. These changes comprise the morphological basis of human speech production which is characterized by sequential articulatory gestures coproduced with the melody of glottal sounds. This suggests that these human specific changes in the body form modify the neural maps of motor and sensory organs in order to facilitate the functional linkage between speech production and perception.

Although the anatomical relationship between the tongue and larynx is free from rigid anatomical coupling, some vestigial connections induce tongue-larynx interactions which are observed both in the articulatory and auditory domains.

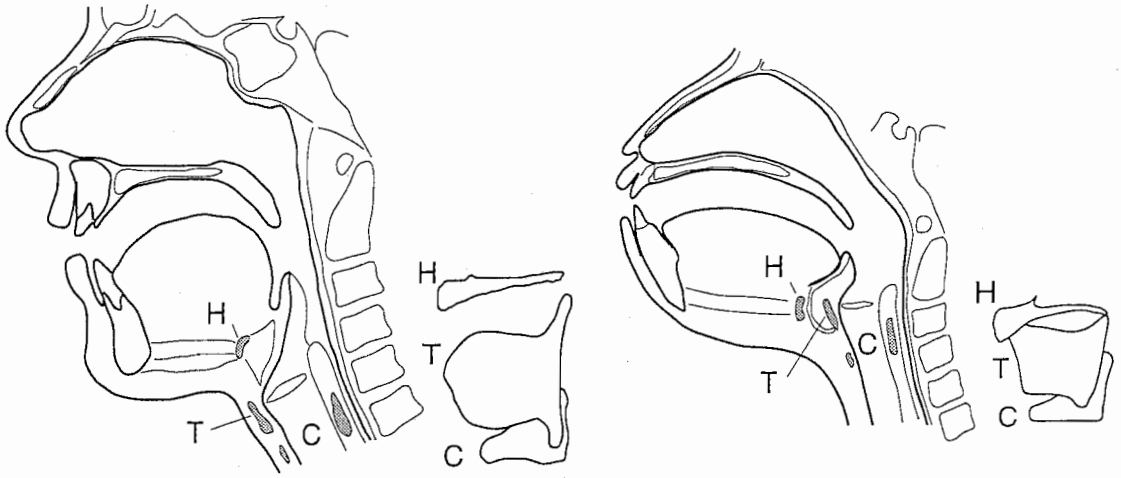


Fig. 1. Comparative morphology of the speech organs in human and macaque. The human vocal tract shape is characterized by the descent of the larynx, the increase of the pharyngeal cavity, and the separation of the thyroid cartilage (T) from the hyoid bone (H). These changes provide a morphological basis of human speech production composed of a large tongue deformation along with continuous vocal fold vibration.

3. Larynx Height and Fundamental Frequency

It is widely supposed that the relative height of the larynx tends to change with F0, being higher for high F0 and lower for low F0. In spite of this will-known empirical observation, there are a couple of problems regarding the relationship between larynx height and F0. The first one is that the tendency is not always seen during natural speech utterances. This is partly due to various articulatory effects on larynx height which deteriorates its relationship with F0. Another account may be that the intended pattern of F0 control in natural speech is the relative change in F0, since the absolute F0 value should vary depending on various prosodic contexts. The second problem is that the mechanism of F0 control by vertical laryngeal movement is not understood despite the obvious phenomenon. The current speech physiology does not provide reasonable explanations on how larynx height affects vocal fold length or tension. The following paragraphs are the summary of our study to answer the above questions. We examined morphological data recorded during sustained phonation of a vowel in different F0 targets in order to explore the relationship between larynx height and F0. In static gestures for sustained production of isolated vowels, the consistent relationship between larynx position and F0 is expected to be seen because vowel quality and tonal specification of utterance have to be realized independent from articulatory and prosodic contexts.

Fig. 2(a) shows an example of our experimental results obtained by the magnetic resonance imaging (MRI) technique (Hirai, Honda, Fujimoto, & Shimada, 1994; Honda, Hirai, & Kusakawa, 1993). A series of MRI scans were performed for each tone separately while the subjects produced a descending scale in the vowel /a/. Larynx positions traced in

the figure indicate a monotonic laryngeal descent during F0 lowering for all the subjects of the experiment. The result of this study suggests a plausible mechanism of F0 control via the positional change of the larynx, as shown in Fig. 2(b). When the whole larynx moves vertically, the posterior plate of the cricoid cartilage shows a sliding motion along the cervical spine. Since the cervical spine has a natural curvature called "lordosis" at the level of the cricoid cartilage, the vertical motion of the larynx automatically causes a rotation of this cartilage. Subsequently, a change in vocal fold length is produced by the same manner that the cricothyroid muscle stretches the vocal folds. From a viewpoint of comparative morphology, the descent of the larynx and the lordosis of the cervical spine are the elements of human specific form of the body. These evolutionary changes do not only contribute to expanding the range of F0 control but also facilitate the sensorimotor mapping between larynx height and tone height.

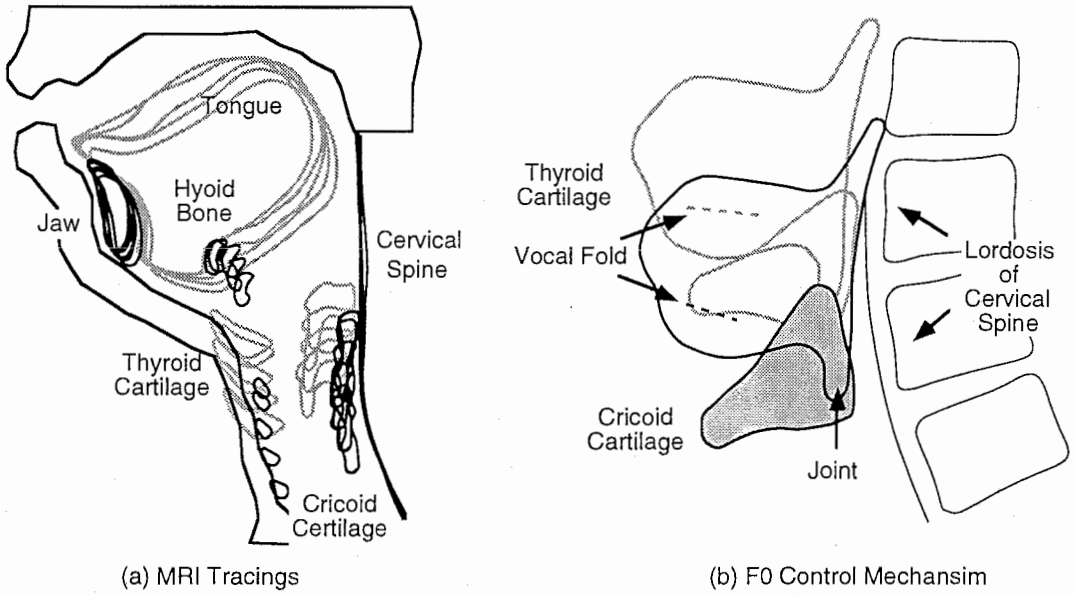


Fig. 2. The mechanism of F0 control by vertical laryngeal movements. (a) The tracings of MRI scan during a musical scale indicate a correlation between larynx position and F0. (b) Vertical laryngeal movements near the maximum point of the cervical lordosis produces a cricoid rotation for varying vocal fold length.

The F0 change in human sound production is accompanied by various events in the body. They include the changes in vocal fold length, cricothyroid angle, larynx height, jaw positions, tongue shape, and subglottal pressure. Each of these events maintains a consistent relation with F0, accounting for a high phenomenological correlation. The motor events are monitored by the brain via various sensory organs. Among them, the muscles that elicit body actions also serve as an integrated sensory system to detect the length, tension, and weight of every part of the body. In speech production, the results of these actions are reflected by the sound, and monitored via the auditory channel as well. In the brain, the generation of a motor plan for F0 change is always followed by somatosensory and auditory information of the produced F0 change. The relationship between the intended motor pattern and received sensory information may be evaluated in the brain just like a correlation analysis. When the motor and sensory patterns have robust analogous relation, they should facilitate the connectivity between the neural map of the body and the auditory image of the sound.

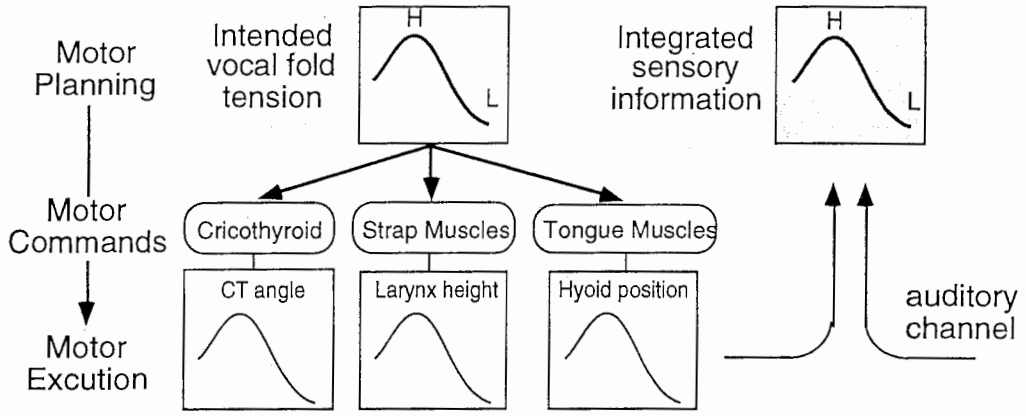


Fig. 3. Various physiological events during F0 changes. The intention of F0 production is reflected by the changes in cricothyroid angle, larynx height, hyoid bone position and the sounds. These events are monitored by the brain to form integrated sensory information of F0 changes. Since these motor and sensory patterns are analogous, they enhance robust sensorimotor mapping.

4. Tongue Shape and Vowel Formants

The similarity between articulatory and auditory patterns has also been acknowledged as a phonetic characteristics of vowels. In the case of vowels, the relationship between the patterns is not linear but multi-dimensional. In the phonetic literature, the vowel system is described by the cardinal vowel chart, which maps the highest point of the tongue in the lateral view. Acoustically, it is defined by a formant diagram, i.e., a plot of the first and the second formant frequencies. It is well-known that the vowel distributions in these kinematic and acoustic spaces resemble each other, as depicted in Fig. 4. Thus, the tongue position for a vowel can be predicted by its acoustic pattern. The analogous relationship between vowel's articulatory and auditory patterns has been discussed in Kojima's study on chimpanzee's vowel perception (1988). He speculates that the coincidence between vowels' auditory and articulatory patterns is a result of human evolution which occurred to the organs of speech production and perception.

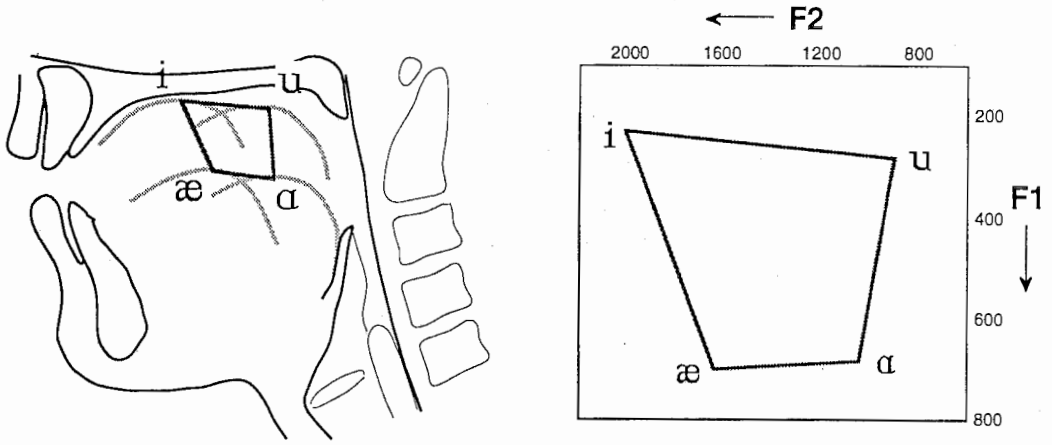
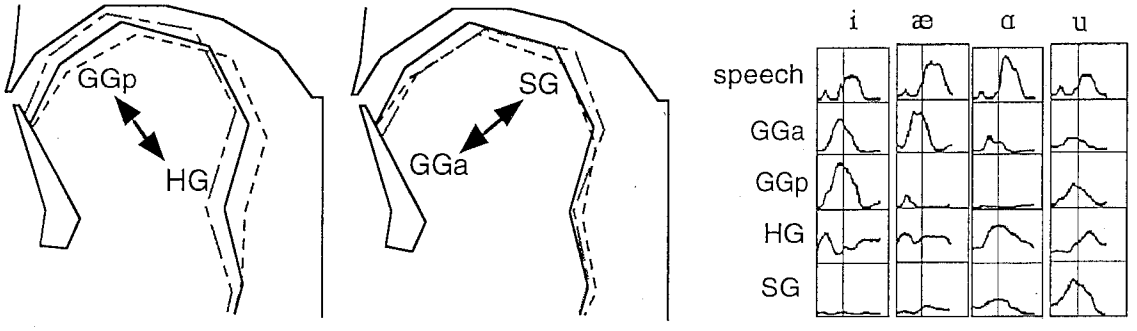


Fig. 4. A schematic drawing of the relationship between vowel's kinematic and acoustic patterns. The distribution of the highest points of the tongue for vowels (left) resembles the acoustic distribution of vowels in the F1-F2 diagram (right).

Our study of tongue muscle function has shown evidence that motor patterns for vowel production also resemble the vowel distribution in the formant space (Kusakawa, Honda, & Kakita, 1993; Honda, Hirai, & Kusakawa, 1993). The electromyographic (EMG) data from the four extrinsic tongue muscles have been reported by Baer, Alfonso, & Honda (1988). We analyzed the data to reconstruct the "intended gesture" for speech utterances by computing the equilibrium point of muscle forces. The human extrinsic tongue muscles are organized to have two pairs of antagonistic muscles. As shown in Fig. 5, the genioglossus posterior (GGp) and the hyoglossus (HG) form a pair, and the styloglossus (SG) and the genioglossus anterior (GGa) another. Since the axes of the muscle pairs are roughly orthogonal, the equilibrium point of muscle forces is computed by a simple vector summation of EMG values for all the muscles. Fig. 6 shows the trajectories of the equilibrium point in the word utterances /əpVp/ having four extreme vowels /i, æ, ɑ, u/. The extreme points in the trajectories show a clear separation of motor targets for vowels which resembles the vowel distribution in the formant space. This result suggests that the neural process of articulatory-to-auditory mapping may be relatively simple. A further study has demonstrated that computational mappings from muscle force equilibrium to formant patterns are successfully achieved by adding the jaw and lip components in the equilibrium equations (Maeda & Honda, 1994). The fact that the neuromotor patterns of vowel production resemble the auditory patterns may explain the efficiency of human speech communication.



(a) two pairs of antaginistc tongue muscles (b) Integrated EMG waveform

Fig. 5. The extrinsic tongue muscles and the EMG data. (a) The four extrinsic tongue muscles are organized as two pairs of antagonists (GGp-HG, and SG-GGp). (b) The EMG data during /əpVp/ utterances with English vowels /i, æ, ɑ, u/.

The coincidence of the auditory and articulatory representations of vowels stems from the shape of the human vocal tract. A vocal tract model of a straight acoustic tube can demonstrate two-dimensional distribution of vowels in the formant space. The synthesized sounds from the model with adequate constriction location can be grouped into vowel categories by the human auditory system. Similar to such a model, tongue deformation in the oral cavity also generates a constriction at various loci along the entire vocal tract. However, the human speech production system produces a quite different effect. Because of the right-angled vocal tract and the orthogonal tongue muscles, they form analogous motor and sensory patterns of vowels. Thus, the shape of a particular part of the body contributes to a robust correspondence between kinematic and acoustic patterns of vowels. The spatial correlation between them is also repeatedly represented in the brain, and it enforces the functional linkage between vowel production and perception.

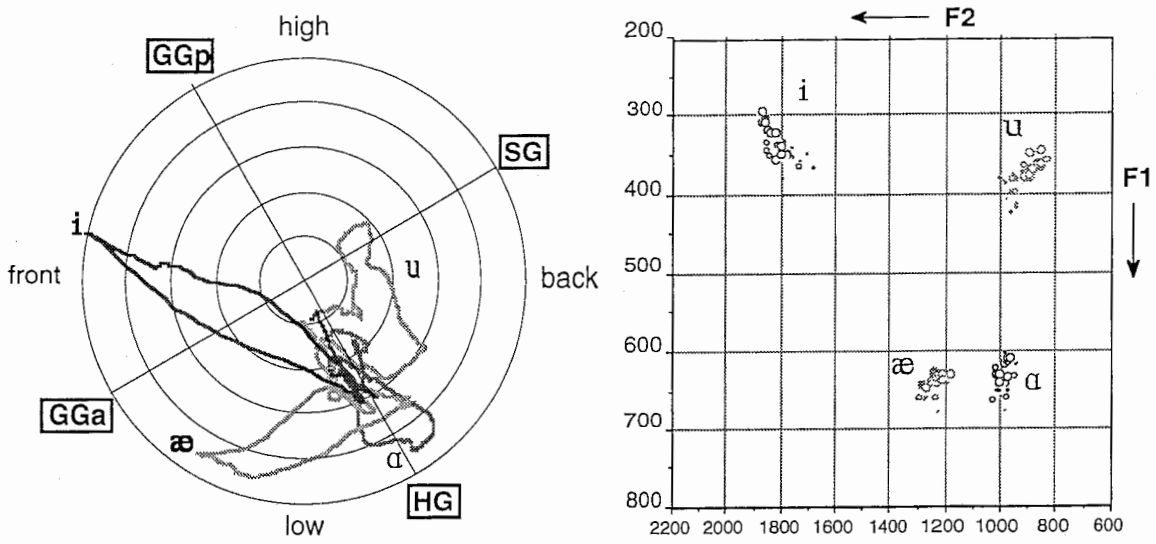


Fig. 6. The EMG and acoustic data from /əpVp/ utterances indicating analogous distribution of vowels in the articulatory and auditory spaces. (a) Articulatory trajectories of the equilibrium point of tongue muscle forces showing vowel distribution in the motor space. (b) The vowel distribution in the formant space.

An additional example of analogous articulatory and auditory patterns of vowels is found in the recent x-ray microbeam study at the University of Wisconsin (Hashi, Westbury, & Honda, 1994). They analyzed kinematic and acoustic data of English and Japanese vowels in order to explore the variability of vowels in the both domains. One of their results indicated that the correspondence between tongue position and auditory pattern of vowels may be invariant across languages. In the kinematic space, English vowels show an evenly distributed pattern, while Japanese vowels demonstrate a clustering of the vowels /i, e, u/. The vowel distribution in the auditory space also demonstrate a similar tendency of vowel clustering for Japanese data, indicating a consistent relationship between vowels' articulatory and auditory distributions. From a linguistic point of view, the vowel system is different across languages, and the pattern of vowel distribution is dependent on the language's phonology. The data described above, however, indicate that the deviation of the vowel's kinematic patterns in a vowel system is also reflected by analogously deformed vowel distribution in the auditory space. According to the idea of the somatoneural relation, it is suggested that the human vocal tract is uniquely formed so that motor representation of vowel production can efficiently induce its auditory representation.

5. Integrated Representation of Vowel and F0

The above observations provide a macroscopic perspective on the idea that the auditory-articulatory linkage derives from the somatoneural relation. The analogous patterns speculated between articulatory and auditory representations of speech components imply a functional formulation of a "speech module" across the motor and the sensory areas in the brain. The speech module conceptualized in this paper consists of the submodules for vowel and F0 as schematically represented in Fig. 7. This scheme is partly based on the cortical

localization of speech function which is well-known as speech areas of Broca and Wernicke. Within each submodules, the articulatory and auditory representations are mapped to each other via the arcuate fasciculus in order to establish bidirectional auditory-articulatory linkage. In contrast to the traditional neuropsychological account, the tight linkage of analogous patterns infers functional equivalence of the motor and sensory representations of speech, enforcing a global account that the two speech areas share the same information. Furthermore, the submodules of vowel and F0 are not independent from each other, but they appear to interact as a functionally integrated unit. This is supported by phonetic and psychoacoustic evidence with respect to vowel and F0.

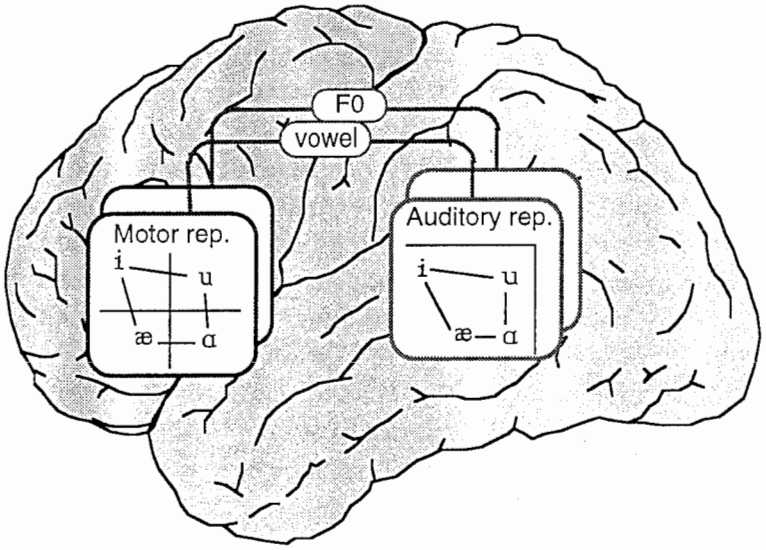


Fig. 6. The speech module of integrated representation of vowel and F0. Vowel and F0 have analogous representations in the articulatory and auditory spaces, which provides efficient sensorimotor coordinate transformation.

The phonetic evidence of vowel-F0 relationship is well demonstrated by the intrinsic vowel F0, which is referred to as the language universal tendency for vowel height and F0 to be correlated. One of the plausible explanations of the intrinsic vowel F0 is the biomechanical interaction between the larynx and the supra-laryngeal articulators (Honda, 1983). The tongue deformation for vowel articulation influences laryngeal configuration via the positional variation of the hyoid bone that interconnects these two organs. In contract to this physiological account, a different explanation in the perceptual domain is also possible, as seen in the "speech enhancement" (Diehl, 1991) or the "auditory dispersion" (Lindblom, 1986) hypotheses. These theories indicate that deliberate production of the intrinsic vowel F0 contributes to robust perceptual separation of vowel quality. These contrasting accounts of the intrinsic vowel F0, however, do not contradict, both allowing one to suggest that the representations of vowel and F0 are unified in the speech module.

Another example of the vowel-F0 integration is the well-known perceptual function of "talker normalization." The perception of vowel quality is not entirely dependent on vowel distribution in the formant space, but is also affected by F0. This perceptual normalization of vowel quality by formants and F0 plays an important role in speech acquisition through mother-infant communication. The fact that the vowels produced by the talker's vocal tracts

of different size are perceived as equivalent in quality by the listener suggests an innate perceptual mechanism which compensates for the difference in the size of the talker's speech organs. It is also known that the vowel normalization is not only auditorily relevant but also involves visual function. Early vision is mainly sensitive to the shape of the object rather than to its size. In visual perception of vowels, the shapes of the lips provide analogous vowel information even if the vocal tracts are different in size.

While these explanations in the perceptual domain appear plausible, there are possible articulatory accounts of talker normalization. The covariance between formants and F0 that enhances the perception of vowel quality is also found in an F0-related articulatory effect, which is termed the "inverse effect" of the intrinsic vowel F0 (Honda, in press). The mechanisms of F0 control do not only involve laryngeal muscles but also tongue and jaw muscles, and the use of these muscles for F0 control inevitably alters the articulatory configuration for vowels. Consequently, a systematic variation of vowel formants results due to F0 control. Although the physiological mechanism of the inverse effect of the intrinsic F0 is complex and has not been examined in detail, the most obvious aspect of the phenomena may be that vowels produced with high F0 tends to have a forward position of the tongue root, as seen in Fig. 2(a). The subsequent acoustic effect of producing high F0 is a higher shift of F2. Thus, the characteristics of the articulatory system produce a tendency for F2 to covary with F0. This acoustic effect of F0 control mechanism on vowel formants is roughly equivalent to the auditory effect that enhances vowel normalization, and it possibly facilitates a "self-emulation" for talker normalization. Although this account from an articulatory view may be premature, it may provide another evidence to support the integrated representation of vowel and F0 in the speech module.

6. Summary

Speech sounds are produced by human vocal tract organs. In a conventional account, action patterns of these organs are controlled by the brain. The concept of the somatoneural relationship emphasizes a contradictory idea that the brain organization must adjust to the shape of the body. The form of the speech organs and the acoustic pattern of the sounds produced by them are represented in the brain so that they are functionally associated. The human-specific shape of speech organs permits the significant correlation between sound patterns and motor patterns for vowel and F0, which allows an auditory-articulatory linkage to form a speech module during the process of neurogenesis.

Speech sounds are received initially in the both sides of the primary auditory cortex in the same manner as other environmental sounds are perceived. However, speech sounds are transmitted to the speech areas in the left side of the brain, where the speech module exists as an innate functional structure. This module is formed to integrate the sensorimotor maps of the body, serving as an interface to the organization of language. The module functions in a way to bridge the motor and sensory association areas for facilitating bi-directional informational flow between them. With this functional linkage, the intention of speaking is monitored as an auditory image prior to its execution. Also, the auditory image of received speech sounds is represented as a corresponding motor pattern.

Vowel and F0 are the core components of the speech module, and they also interact with each other. The interaction is observed as the intrinsic vowel F0 and vowel normalization by F0. They have been discussed as different issues, however they can be summarized by the same concept of the vowel-F0 integration in the speech module. The integrated image is transparent in the representations of speech production and perception. This proposes a resonance hypothesis for the human brain function to fuse the images of sound and action.

References

Baer, T., Alfonso, P. J., & Honda, K. (1988). Electromyography of the tongue muscles during vowels in /əpVp/ environment. *Annual Bulletin of Research Institutue of Logopedics and Phoniatrics,* 22, 7-19.

Diehl, R. L. (1991) The role of phonetics within the study of language. Phonetica, 48, 120-134.

Hashi, M., Westbury, J. R., & Honda, K. (1994). Articulatory and acoustic variability of vowels in Japanese and English. *Journal of Acoustical Society of America,* 95, Pt. 2, 2820.

Hirai, H., Honda, K., Fujimoto, I., & Shimada, Y. (1994). Analysis of magnetic resonance images on the physiological mechanisms of fundamental frequency control. *Journal of Acoustical Society of Japan*, 50, 296-304. (in Japanese)

Honda, K. (1983). Relationship between pitch control and vowel articulation. In D. M. Bless, & J. H. Abbs (eds.), *Vocal Fold Physiology* (pp. 286-297), San Diego: College-Hill Press.

Honda, K., Hirai, H., & Kusakawa, N. (1993). Modeling vocal tract organs based on MRI and EMG observations and its implication on brain function. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics,* 27, 37-49.

Honda, K. (in press). Laryngeal and extra-laryngeal mechanism of F0 control. In F. Bell-Berti, & L. J. Raphael (eds.), *The Festschrift for K. S. Harris.*

Kojima, S. (1988). Audition, speech perception and phonation of the chimpanzee: a search for the origin of human speech. Primate Res. 4, 44-65. (in Japanese)

Kusakawa, N., Honda, K., & Kakita, Y. (1993). Construction of articulatory trajectories in the space of tongue muscle contraction force. *ATR Technical Report, TR-A-0171.*

Lindblom, B. (1986) Phonetic universals in vowel systems. In J. J. Ohala, & J. J. Jaeger (eds.), Experimental Phonology, Orlando, Academic Press, pp.13-44.

Maeda, S., & Honda, K. (1994). From EMG to formant patterns of vowels: the implication of vowel spaces. Phonetica, 51, 17-29.

Purves, D. (1988). *Body and Brain: a Trophic Theory of Neural Connections.* Cambridge, Mass: Harvard University Press.

The conceptual basis of modelling auditory processing in the brainstem

Ray Meddis,

Speech and Hearing Laboratory, University of Technology
Loughborough, U.K.

Computer modelling of the neurophysiology of the auditory brainstem has the potential for providing a coherent framework for synthesising the growing knowledge database generated by anatomists and physiologists. It can also supply hypotheses concerning the mechanisms underlying psycho acoustic phenomena. The framework should, furthermore, help us to enumerate the underlying principles of sensory analysis independently of the wetware or hardware which embody them. These principles can also be used selectively by engineers when solving auditory signal processing problems.

If this effort is to develop into a mature science, we need to identify and make explicit the basic conceptual building blocks which support the modelling process. Similarly, if we are to interest engineers in the potential of incorporating simulations of living systems into their devices we must be able to identify the functions and benefits of the individual components. They will not be willing to slavishly copy the whole auditory system but may well wish to exploit the power of individual principles if we can isolate them and articulate them.

For the purpose of this exposition, I shall take the view that the acoustic signal carries information which could influence the action of an animal. The function of the information processing system is to amplify this information at the expense of other aspects of the signal. A secondary function of the system is to segregate different types of information so that each can be separately amplified and directed into the correct action-channel. Below, I shall distinguish three quite different aspects of the signal processing used by mammals to achieve this selectivity and amplification. Firstly, purely physical effects prior to nervous processing. Secondly, variable types of response by individual neuronal components and, thirdly, patterns of responding across groups of neurones. In the talk I shall illustrate some of these principles using recent modelling work at Loughborough.

Physical effects

The most approachable and best understood set of transforms can be observed before the signal even reaches the auditory meatus. The 20 cm separation of the two ears, the interposition of the head between the two ears and the complex convolutions of the pinna's reflecting surfaces combine to ensure that a single sound source gives rise to two quite different acoustic waveforms at the entrance to the left and right meatus. The difference in the two signals contains a great deal of information concerning the location

of the sound source which could not have been discovered using the original signal alone. Moreover, the combined use of the left and right signals can later be used to increase the detectability of a signal against a noisy background (BMLD effect). The ability of the system to orient the head appropriately further enhances these effects. This actively controlled creation of two different signals from the same source should be regarded as the first stage in the information processing chain and should ideally feature prominently in any complete model of hearing.

The resonance of the concha and the meatus gives rise to the second signal processing stage by selectively amplifying signals in a broad but limited frequency range. In man, the concha and the meatus combine to provide 10 dB amplification to signals between 1 kHz and 7 kHz (Shaw, 1974). Curiously, the 'head shadow' effect also amplifies signals of contralateral origin in the mid frequency range while attenuating those at higher frequencies. Whether our speech frequency range is tuned to take advantage of the length of our meatus, or the length of the meatus has been adapted to suit the speech is a moot point but it does constitute a clear functional compatibility.

The frequency selectivity in the cochlea is a clear example of separation of different aspects of the signal prior to further processing. A great deal of current auditory theorising is currently concerned with the use that this is put to later in the system. However, there is more to this stage than mere filtering. The nonlinearity of the basilar membrane response introduces a range of effects whose functional significance is only just beginning to be explored. At low amplitudes, filters are narrow but at high amplitudes they are wide. This is the opposite of what one might expect if the intention were to optimise the detectability of weak signals. So what is the purpose of this arrangement?

The nonlinear response also generates distortion products which convert simple inputs into complex outputs. A hi-fi engineer is concerned to eliminate such products, so why would mammalian hearing introduce them? One possibility is that they enrich the signal and distribute it across a larger number of filters. This may enhance its delectability in some way. In the case of harmonic sounds, the distortion products are all related harmonically to the fundamental frequency and this may emphasise periodic aspects of the signal. The most striking effect of nonlinearity is two-tone suppression which has the potentially useful effect of causing strong signals to suppressing neighbouring weaker signals. This should, for example, emphasise formant peaks in speech signals.

With the exception of filterbanks, the potential of these other physical signal processing principles have not been vigorously explored by engineers. Nor would we expect them to do so until the hearing community has established a consensus of the functional importance of each stage.

Individual neuronal components

The inner hair cell (not strictly neuronal) is the most familiar component to auditory modellers (Hewitt and Meddis, 1991, for review). Its complex

response to basilar membrane vibrations introduces a number of interesting transformations to the signal. It is directly responsible for signal adaptation, low-frequency phase-locking and the probabilistic response of the auditory nerve fibre. We can readily market the benefits of adaptation (more easily spotted signal onsets), of phase-locking (carries information about the signal fine-structure) and probabilistic response (distributes the information efficiently across a number of fibres each with limited information transmission capabilities). Unfortunately the explanations are incomplete. Adaptation only occurs for intense signals and is not present near threshold and phase-locking is restricted to low frequencies in many (but not all) animals. When inner hair cell models are added to automatic speech processing devices, they more often than not produce a reduction in performance. Clearly its functional benefits need to be better understood before they can be exploited in signal processing devices.

Most models ignore the probabilistic response of individual auditory nerve fibres even though this may be the main (possible the only) substantial source of system noise in the auditory nervous system. Its inclusion is probably essential when modelling sensory thresholds. The number of fibres is strictly limited and the common assumption that groups of fibres carry an essentially noise-free representation is probably not warranted given the small numbers innervating individual frequency regions.

Refractory effects are also typically regarded as a nuisance factor. Certainly, any pulse code modulated system must have gaps to define the beginning and end of the pulses. But is there more to it than that? We were surprised to discover that the refractory period of stellate cells in the cochlear nucleus may be the prime determinant of the cell's chopping rate. That, in turn, determined the characteristics of its bandpass amplitude modulation transfer function. The recovery rate of the cell was used as a critical component in the cell's ability to selectively amplify certain rates of amplitude modulation whilst attenuating others (Hewitt et al, 1992). Different cells with different rates of recovery consequently amplify modulation in different frequency regions - an essential step in the process of mapping signal periodicity onto a place code.

The style of recovery of a cell to a disturbance of its equilibrium can also have an effect on how it process information. When a cell with a linear input/output current/voltage response function has its internal voltage raised following synaptic input, K+ currents resist this rise and a Na spike is generated only if the input is prolonged and substantial. Stellate cells are like this and act as integrators for many inputs. Each input is subthreshold but has a lingering effect on the cell so that inputs can accumulate over time to generate an adequate driving stimulus.

Other types of cells respond to the rising voltage by applying an accelerator function which results in a very early Na spike. Bushy cells fall into this category and they consequently have a very fast response to one or a small number of inputs. When the response requires more than a single input, these must arrive virtually simultaneously because the fast time-constant of the cell means that the effect of a single subthreshold input does not linger long enough to interact with later inputs.

The fusiform cell in the DCN represents yet another type of response. A brief period of hyper polarisation caused by inhibition triggers current flow which will continue to resist excitatory inputs even after the inhibition has ceased. This has been used to explain the 'build up' pattern of responding which is typical of these cells. The functional benefit of this type of response is still unclear, however.

The number of excitatory synapses tolerated by a cell can also determine the style of response of that cell. For example, we have shown that a sustained chopper response pattern can be converted to a transient chopper pattern by reducing the number of AN synapses while keeping the threshold constant (Hewitt and Meddis, 1993). The AM amplification properties of the cell will deteriorate under these circumstances. It has also been suggested that an onset pattern of responding can be simulated by using a cell with a large number of AN inputs and a high threshold. Such cells will only respond at the onset of a stimulus when most of the input fibres have a high probability of firing.

Patterns of response across neurones

A fundamental aspect of patterned response is the geographical segregation of neurones on the basis of the information extracted from the signal. The auditory nervous system contains many 'maps', including frequency maps, source location maps, and amplitude modulation frequency maps. The significance of maps may extend beyond the simple principle of keeping different properties separate; they may also serve the purpose of keeping cells processing similar aspect of information together so that they can interact meaningfully. The most obvious interaction is that of lateral inhibition where peaks of activity across a dimension can be highlighted and emphasised.

Other co-operative circuits are slowly being identified. For example, it has been suggested that the complex circuitry of the DCN allows fusiform cells to respond optimally to notches in broadband noise. An 'echo-suppression' circuit has also been proposed involving vertical cells in the DCN that deliver a slightly delayed inhibition to VCN bushy cells (Wickesberg and Oertel; 1993). These would suppress a second burst of AN activity following a reverberation-induced echo. To these we should add circuits in the MSO and LSO which oppose inputs from opposite sides to extract information about the location of sound sources.

Feedback circuits from the superior olive to bushy cells in the AVCN provide on-centre inhibition which reduces overall firing rate and may enhance the cell's dynamic range (Caspary et al; 1993). The MOC system sends descending projections to the auditory periphery where it can raise response thresholds. Intriguingly, these connections also send collaterals to the CN where they may impact on the same cells that receive input from the periphery. This circuit may be a complex feedback arrangement where the system is allowed to anticipate the effects of the longer feedback loop and thus reduce the likelihood that the system will overshoot or hunt; both problems with simple feedback systems.

Conclusions

We are clearly still on the threshold of understanding how brainstem circuits are used to process auditory information. Even the more familiar territory of the auditory periphery still has unanswered questions concerning the functional benefits of the individual process that are observed there. A convergence of information from psychophysics, physiology and anatomy is providing a rich database from which to speculate and computer modelling will be the testbed of these theories and the embodiment of the synthesis which results.

References

[1] Caspary, D.M., Palumbi, P.S. Backoff,P.M. Helfert,R.H. and Finlayson,P.G. (1993) " GABA and glycine inputs control discharge rate within the excitatory response area of primary-like and phase-locked AVCN neurons" In The Mammalian Cochlear Nuclei: Organization and Function, Merchan, M.A. et al, Plenum.

[2] Hewitt J.Michael., Meddis Ray. (1993) " Regularity of cochlear nucleus stellate cells: A computational modeling study" J.Acoust. Soc. Am. 93 (6) 3390-3399

[3] Hewitt Michael J., Meddis Ray.,, Shackleton Trevor M.(1992) " A computer model of a cochlear-nucleus stellate cell: Responses to amplitude-modulated and pure-tone stimuli" J.Acoust. Soc. Am. 91(4)

[4] Hewitt Michal J., Meddis Ray (1991) " An evaluation of eight computer models of mammalian inner hair-cell function" Journal Acoustical Society America 90, 904-917

[5] Shaw, E.A. (1974) the External Ear. In Handbook of Sensory Physiology, (W.D.Keidel and W.D. Neff) Vol 5/1 pp 455-490. Springer Berlin.

[6] Wickesberg,R.E. and Oertel,D.(1993) " Intrinsic connections in the cochlear nucleus complex studied in vitro and in vivo" In The Mammalian Cochlear Nuclei: Organization and Function, Merchan, M.A. et al, Plenum.

ROBUST SPEECH RECOGNITION BASED ON HUMAN BINAURAL PERCEPTION

Richard M. Stern and Thomas M. Sullivan

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

In this paper we present a new method of signal processing for robust speech recognition using multiple microphones. The method, based on human binaural hearing, consists of passing the speech signals detected by multiple microphones through bandpass filtering and nonlinear rectification operations, and then cross-correlating the outputs from each channel within each frequency band. These operations provide an estimate of the energy contained in the speech signal in each frequency band, and provides rejection of off-axis jamming noise sources. We demonstrate that this method increases recognition accuracy for a multi-channel signal compared to equivalent processing of a monaural signal, and compared to processing using simple delay-and-sum beamforming.

1. INTRODUCTION

The need for speech recognition systems and spoken language systems to be robust with respect to their acoustical environment has become more widely appreciated in recent years. Results of several studies have demonstrated that even automatic speech recognition systems that are designed to be speaker independent can perform very poorly when they are tested using a different type of microphone or acoustical environment from the one with which they were trained, even in a relatively quiet office environment (*e.g.* [1]). Applications such as speech recognition over telephones, in automobiles, on a factory floor, or outdoors demand an even greater degree of environmental robustness.

In recent years there has been increased interest in the application of knowledge about signal processing in the human auditory system to improve the performance of automatic speech recognition systems (*e.g.* [2, 3, 4]). With some exceptions (*e.g.* [5, 6]), these algorithms have been primarily concerned with signal processing in the auditory periphery, typically at the level of individual fibers of the auditory nerve. While the human binaural system is primarily known for its ability to identify the locations of sound sources, it can also significantly improve the intelligibility of sound, particularly in reverberant environments [7]. In this paper we describe an algorithm that combines the outputs of multiple microphones to improve speech recognition accuracy. The form of this algorithm is motivated by knowledge of the more central processing that takes place in the human binaural system.

Since our algorithm processes the outputs of multiple microphones, it should be evaluated in comparison with other microphone-array approaches. Several types of array processing strategies have been applied to speech recognition systems. The simplest such system is the delay-and-sum beamformer (*e.g.* [8]). In delay-and-sum systems, steering delays are applied at the outputs of the microphones to compensate for arrival time differences between microphones to a desired signal, reinforcing the desired signal over other signals present. This approach works reasonably well, but a relatively large number of microphones is needed for large processing gains. A second approach is to use an adaptive algorithm based on minimizing mean square energy, such as the Frost or the Griffiths-Jim algorithm [9]. These algorithms can provide nulls in the direction of undesired noise sources, as well as greater sensitivity in the direction of the desired signal, but they assume that the desired signal is statistically independent of all sources of degradation. Consequently, they do not perform well in environments when the distortion is at least in part a delayed version of the desired speech signal as is the case in many typical reverberant rooms (*e.g.* [10]). (This problem can be avoided by only adapting during non-speech segments [11].)

The algorithm described in this paper is based on a third type of processing, the cross-correlation-based processing in the human binaural system. The human auditory system is a remarkably robust recognition system for speech in a wide range of environmental conditions, and other signal processing schemes have been proposed that are based on human binaural hearing (*e.g.* [12]). Nevertheless, most previous studies have used cross-correlation-based processing to identify the direction of a desired sound source, rather than to improve the quality of input for speech recognition (*e.g.* [14, 15]).

In Sec. 2 we briefly review some aspects of human binaural processing, and we describe the new cross-correlation-based algorithm in Sec. 3. In Sec. 4 we describe typical results from pilot evaluations of the cross-correlation-based algorithm that demonstrate the algorithm's ability to preserve spectral contours. Finally, we describe in Sec. 5 the results of a small number of experiments that compare the speech recognition accuracy obtained with the new cross-correlation-based algorithm on conventional delay-and-sum beamforming.

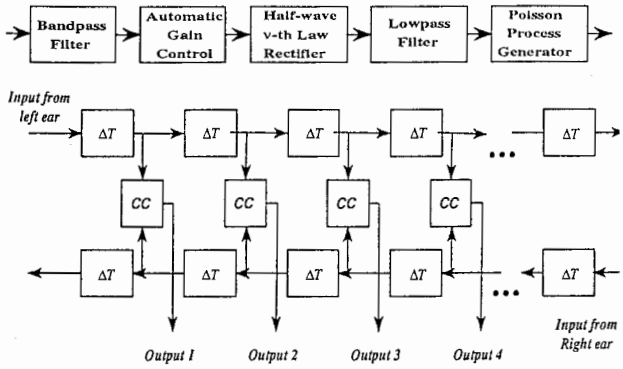


Figure 1. Upper panel: Block diagram of the transduction process in the auditory periphery. The output represents the response of a single fiber of the auditory nerve. Lower panel: Schematic representation of the Jeffress place mechanism. The blocks labelled ΔT indicate fixed timed delays in the signals.

2. CROSS-CORRELATION AND HUMAN BINAURAL PROCESSING

As a crude approximation, the peripheral auditory system can be characterized as a bank of bandpass filters, followed by some nonlinear post-processing. To the extent that such an offhand characterization is valid, we may further suggest that binaural interaction can be characterized as the cross-correlation from ear to ear of the outputs of peripheral channels with matching center frequencies [13].

Figure 1 is a schematic diagram of a popular mechanism that can accomplish the interaural cross-correlation operation in a physiologically-plausible fashion. This approach was originally proposed by Jeffress [14] and later quantified by Colburn [15] and others. The upper panel of Fig. 1 describes a functional model of auditory-nerve activity. This auditory-nerve model consists of (1) a bandpass filter to represent the frequency analysis performed by the auditory periphery, (2) a rectifier that represents nonlinearities in the transduction process, (3) a lowpass filter that represents the loss of synchrony of the auditory-nerve response to stimulus fine structure above about 1500 Hz, and (4) a mechanism that generates sample functions of a non-homogeneous Poisson process with an instantaneous rate that is proportional to the output of the rectifier.

The lower panel describes a network that performs temporal comparisons of the Poisson pulses arriving from peripheral auditory nerve fibers of the same characteristic frequency (CF), one from each ear, with successive delays of ΔT introduced along the path, as shown. The blocks labelled *CC* record coincidences of neural activity from the two ears (after the net delay incurred by the signals from the peripheral channels by the ΔT blocks). The response of a number of such units, plotted as a function of the net internal interaural delay can be thought of as an approximation to the interaural cross-correlation function of the sound impinging on the ear after the bandpass filtering, rectification, and lowpass filtering is performed by the auditory periphery. Figure 2 displays the relative amount of activity produced by an ensemble of coincidence-

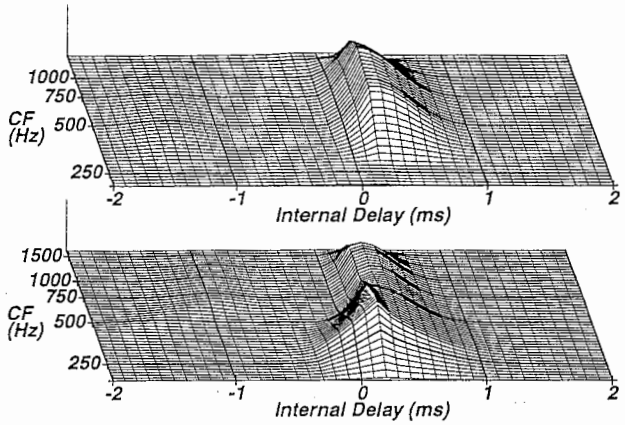


Figure 2. The response of an ensemble of binaural fiber pairs to a 500-Hz pure tone (upper panel) and to bandpass noise centered at 500 Hz (lower panel), each presented with a 0.5-ms ITD.

counting units in response to two simple stimuli: a 500-Hz pure tone, and bandpass noise centered at 500 Hz, each presented with a 0.5- ms interaural time delay (ITD). The expected total number of coincidences is plotted as a function of internal delay (along the horizontal axis) and characteristic frequency (which is represented by the oblique axis). The diminished response for net internal delays greater than 1 ms in magnitude reflects the fact that only a small number of coincidence-counting units are believed to exist with those delays. In traditional binaural models, the location of the ridge along the internal-delay axis is used to estimate the lateral position or azimuth of a sound source. In this work we consider the spectral profile along the ridge (for more complex speech stimuli), and we specifically seek to determine the extent to which the cross-correlation processing of the binaural system serves to preserve the spectral contour along that ridge in difficult environments.

3. CROSS-CORRELATION-BASED MULTI-MICROPHONE PROCESSING

The goal of our multi-microphone processing is to provide a simplified computational realization of elements of the auditory system and of binaural analysis, but with potentially more than two sensors. In other words, we speculate what auditory processing might be like if we had 4, 8, or more ears. Figure 3 is a simplified block diagram of our multi-microphone correlation-based processing system. The input signals $x_k[n]$ are first delayed in order to compensate for differences in the acoustical path length of the desired speech signal to each microphone. (This is the same processing performed by the conventional delay-and-sum beamformer.) The signals from each microphone are passed through a bank of bandpass filters with different center frequencies, passed through nonlinear rectifiers, and the outputs of the rectifiers at each frequency are correlated. (The correlator outputs correspond to outputs of the coincidence counters at the internal delays of the "ridges" in Fig. 2.) Currently we use the 40-channel filterbank proposed by Seneff [2], which was designed to approximate the frequency selectivity of the auditory system. The shape of the rectifier has a significant effect on the results. We have examined

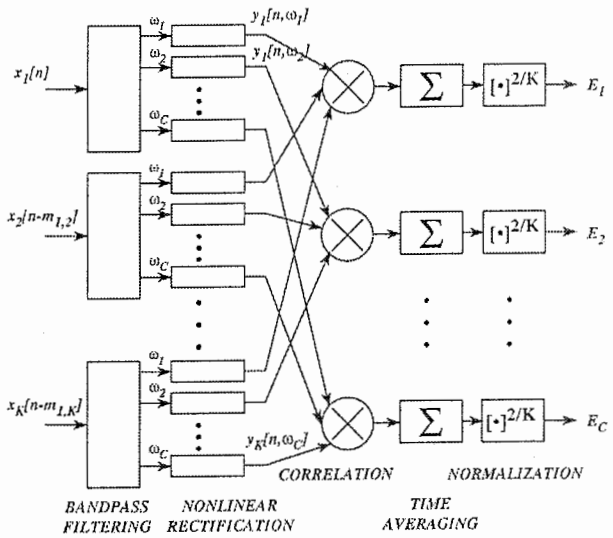


Figure 3. Block diagram of multi-microphone cross-correlation-based processing system.

the response of two types of nonlinear rectifiers: the rectifier originally described by Seneff, which saturates in its response to high-level stimuli, and a family of rectifiers called half-wave power-law rectifiers which produce zero output for negative signals and raise positive signals to an integer power.

For two microphones, these operations correspond to the familiar short-time cross-correlation operation for an arbitrary bandpass channel with center frequency ω_c:

$$E_c = \sum_{n=0}^{N-1} y_1 [n, \omega_c] y_2 [n, \omega_c]$$

where $y_k [n, \omega_c]$ is the signal from the k^{th} microphone after delay, bandpass filtering, and rectification, n is the time index, and N is the number of samples per analysis frame. For the general case of K microphones, these operations produce

$$\hat{E}_c = \{ \sum_{n=0}^{N-1} y_1 [n, \omega_c] \prod_{k=2}^{K} y_k [n, \omega_c] \}^{2/K}$$

The factor of $2/K$ in the exponent enables the result to retain the dimension of energy, regardless of the number of microphones.

The 40 "energy" values are then converted into 12 cepstral coefficients using the cosine transform. The 12 cepstral parameters and an additional coefficient representing the power of the signal during the analysis frame are used as phonetic features for the original CMU SPHINX-I recognition system [16].

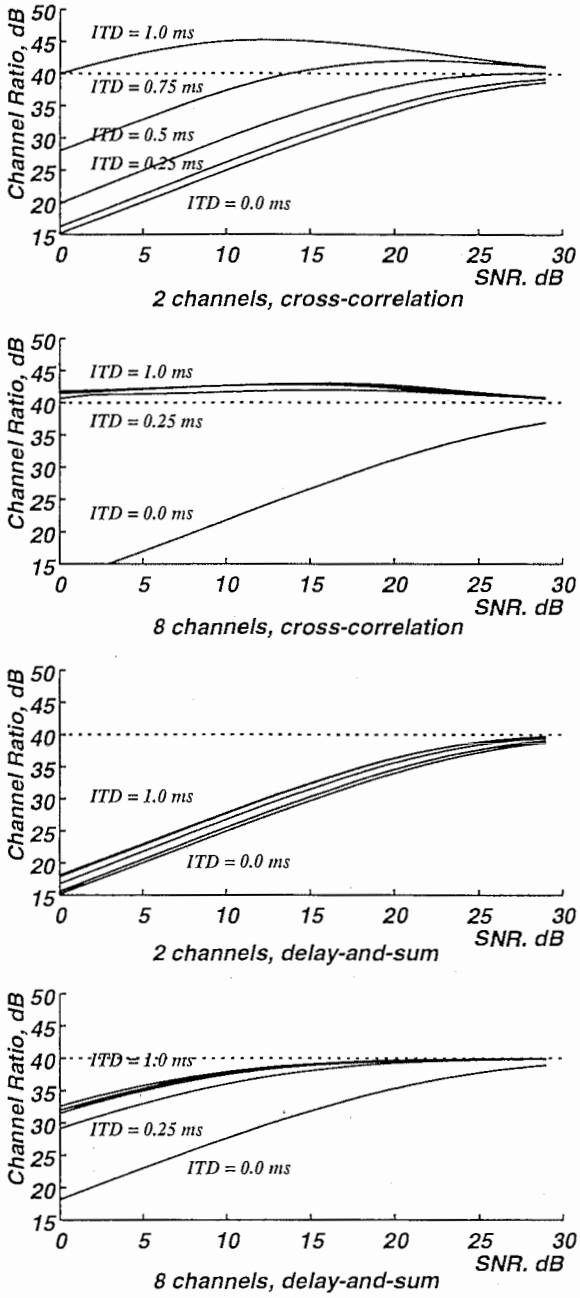


Figure 4. Comparisons of output energies of a 2-channel cross-correlation processor with delay-and-sum beamforming, using artificial additive noise. The actual power ratio of the two tones (without the noise) is 40 dB.

4. CROSS-CORRELATION PROCESSING AND ROBUST SPECTRAL PROFILES

Comparisons using pairs of tones. We first evaluated the cross-correlation algorithm by implementing a series of pilot experiments with artificial stimuli. In the first experiment we examined the spectral profile developed by two sine tones, one at 1 kHz and one at 500 Hz, with an amplitude ratio of 40 dB. The two tones

were summed and corrupted by additive white Gaussian noise. The summed tones were presented identically to each "sensor" of the system (thus representing an "on-axis" signal), but the noise was added with a time delay from sensor to sensor that simulates the delay that is produced when the noise arrives at an oblique angle to a linear microphone array. Each sensor output was then passed through a pair of bandpass filters, one centered at 500 Hz and one at 1 kHz. The signals at the outputs of the bandpass filters were half-wave rectified, and the outputs from filters at corresponding frequency bands from each sensor were cross-correlated to extract an energy value for that frequency band. The ratio of these outputs was calculated and plotted for peak-signal-to-additive-noise ratios (SNR) ranging from 0 dB to 30 dB.

The results of this experiment are depicted in the four panels of Fig. 4, which display the power ratio of the outputs of the 500-Hz and 1000-Hz processing bands, as a function of SNR. In all cases, the ideal result would be the input power ratio of 40 dB, which is indicated by the horizontal dotted lines. Data were obtained for five values of sensor-to-sensor time delay (denoted "ITD"): 0.0, 0.25, 0.5, 0.75, and 1.0 ms. We compare results obtained using the cross-correlation array post processing as described above with processing in which the channels are summed prior to bandpass filtering. This case is representative of delay-and-sum beamforming, where the on-axis sine tone signal is reinforced relative to the off-axis uncorrelated noise signal. It can be seen in Fig. 4 that 8 sensors provides a better approximation than 2 sensors to the original 40-dB ratio of energies in the two frequency channels. For a given number of sensors, the cross-correlation algorithm performs better than delay-and-sum beamforming. Finally, with the desired signals presented simultaneously to the sensors, performance improves (unsurprisingly) as the sensor-to-sensor ITD of the noise is increased.

Comparisons using a synthetic vowel sound. We subsequently confirmed the validity of the algorithm by an analysis of a digitized vowel segment /a/ corrupted by artificially-added white Gaussian noise at global SNRs of 0 to +21 dB. The speech segment was presented to all microphone channels identically (to simulate a desired signal arriving on axis) and the noise was presented with linearly increasing delays to the channels (again, to simulate an off-axis corrupting signal impinging on a linear microphone array). We simulated the processing of such a system using 2 and 8 microphone channels, and time delays for the masking noise of 0 and 0. 125 ms to successive channels.

Figure 5 describes the effect of SNR, the number of processing channels, and the delay of the noise on the spectral profiles of the vowel segment. The frequency representation for the vowel segment is shown along the horizontal axis. (These responses are warped in frequency according to the nonlinear spacing of the auditory filters.) The SNR was varied from 0 to +21 dB in 3-dB steps, as indicated. The upper panel summarizes the results that are obtained using 2 channels with the noise presented with zero delay from channel to channel (which would be the case if the speech and noise signals arrive from the same direction). Note that the shape of the vowel, which is clearly defined at high SNRs, becomes almost indistinct at the lower SNRs. The center and lower panels show the results of processing with 2 and 8 micro-

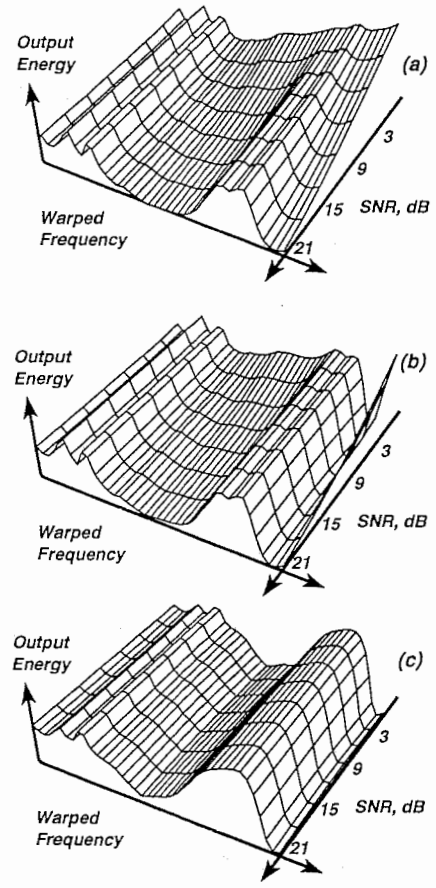


Figure 5. Estimates of spectra for the vowel segment /a/ for various SNR using (a) 2 input channels and zero delay, (b) 2 input channels and 125-μs delay to successive channels, and (c) 8 input channels and 125-μs delay.

phones, respectively, when the noise is presented with a delay of 125 μs from channel to channel (which corresponds to a moderately off-axis source location for typical microphone spacing). We note that as the number of channels increases from 2 to 8, the shape of the vowel segment in Figure 2 becomes much more invariant to the amount of noise present. In general, we found in our pilot experiments that the benefit to be expected from processing increases sharply as the number of microphone channels is increased. We also observed (unsurprisingly) that the degree of improvement increases as the simulated directional disparity between the desired speech signal and the masker increases. We conclude from these pilot experiments that the cross-correlation method described can provide very good robustness to off-axis additive noise. As the number of microphone channels increases, the system is robust to noise at smaller time delays between microphones, so even undesired signals that are slightly off-axis can be rejected.

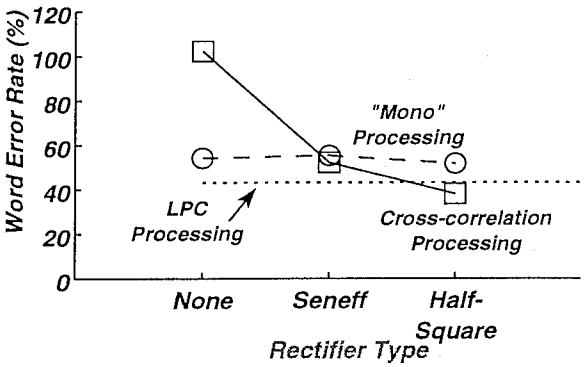


Figure 6. Comparison of word error rates achieved with 2-microphone processing using various half-wave rectifiers, and three types of signal processing.

5. EFFECTS OF CROSS-CORRELATION PROCESSING ON SPEECH RECOGNITION ACCURACY

Encouraged by the appearance of these spectral profiles with simulated input, we evaluated 1-, 2-, 4-, and 8-channel implementations of the algorithm in the context of an actual speech recognition system. The CMU SPHINX-I speech recognizer [16] was trained using speech recorded in an office environment using the speaker-independent alphanumeric census database [1] with the omnidirectional desktop Crown PZM6FS microphone. Identical samples of 1018 training utterances from this database from 74 speakers were presented to the inputs of the multi-microphone system described in Figure 2. All speech was sampled at 16 kHz. The frame size for analysis was 20 ms (320 samples) and frames were analyzed every 10 ms.

5.1. Nonlinear Rectification

The goal of the first series of experiments using actual speech input to the system was to determine the effect of rectifier shape on speech recognition accuracy. A test database was collected using a stereo pair of PZM6FS microphones placed under the monitor of a NeXT workstation. The database consisted of 10 male speakers each uttering 14 alphanumeric census utterances that were similar to those in the training data.

We compared the word errors obtained (tabulated according to the standard ARPA metric) using a 2-channel implementation of the cross-correlation algorithm and a "mono" implementation of the same algorithm in which the same signal is input to the two channels. (The "mono" implementation enables us to assess the extent to which the system can exploit differences between the signals arriving at the two microphones.) We tested with half-wave power-law rectifiers with various exponents, and with the rectifier proposed by Seneff [9]. Figure 4 summarizes the results of these comparisons. Using the half-wave power-law rectifier with the positive signal raised to the 2^{nd} power (the "half-square" rectifier) provided the lowest word error rate of the various half-wave power-law rectifiers. The 2-channel cross-correlation algorithm provides a slightly better error rate than conventional LPC signal processing, and the recognition accuracy using this algorithm depends on the shape of the rectifier.

We hypothesize that the half-square rectifier provides the best error rate because it is slightly expansive. The Seneff rectifier actually compresses the positive signals and limits dynamic range. Using a power-law rectifier of too great a power starts to diminish in performance as the dynamic range is expanded too greatly. Using no rectifier at all provides poor performance because negative correlation values are produced. The half-wave square-law rectifier was used for all subsequent experiments.

5.2. Number of Processing Channels

We describe in this section results obtained using a new set of multiple-channel speech data. This testing database consisted of utterances from the CMU alphanumeric census task [1], and it was collected in a much more difficult environment with significant reverberation and additive noise sources. The ambient noise level was approximately 60 dB SPL with linear frequency weighting. Simultaneous speech samples from a single male speaker were collected using an 8-element linear array of inexpensive noise-cancelling pressure gradient electret condenser microphones, spaced 7 cm from one another. For comparison purposes, each utterance was also simultaneously recorded by a pair of omnidirectional desktop Crown PZM6FS microphones, also spaced 7 cm from one another, and the ARPA-standard Sennheiser HMD-414 close-talking microphone. The subject wore the closetalking microphone and sat at a 1-meter distance from the other microphones. The signals from the electret microphones were passed through a filter with a response of –6 dB/octave between 125 Hz and 2 kHz, and a gain of 24 dB, to compensate for the frequency response of these microphones. By selecting a single element, the middle two elements, or the middle four elements from the 8-element array, arrays of 1, 2, 4, and 8 elements could easily be obtained.

The training database for these experiments was from the original census data, obtained with a PZM6FS microphone with very different acoustical ambience. In order to compensate partially for differences between the training and environments, we normalized each cepstral coefficient (except for the zeroth) on an utterance-by-utterance basis by subtracting the mean of the values of that coefficient across all frames of the utterance.

Figure 7 shows the word error rates obtained using cross-correlation processing with 1, 2, 4, and 8 channels (microphones). The performance of three different algorithms is compared: (1) the original algorithm with auditory processing and the cross-correlation analysis (as in Fig. 3), (2) auditory processing used in conjunction with the initial delay-and-sum beamforming only, and (3) conventional LPC analysis in conjunction with simple delay-and-sum beamforming. It is seen in each case that as more microphones are used, the word error rate decreases. The cross-correlation processing provides lower error rates for the 2- and 4-microphone cases, but all 3 methods give roughly the same performance for the 8-microphone case.

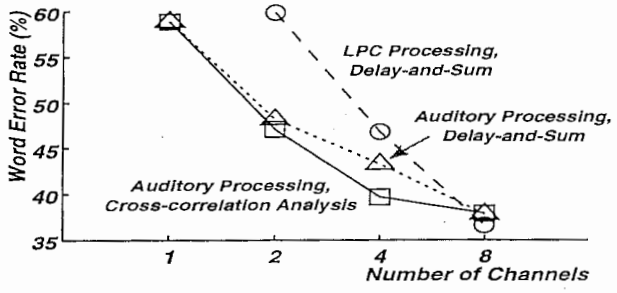


Figure 7. Comparison of word error rates for 1, 2, 4, and 8-channel array processors using the electret microphones of the Flanagan array. The system was trained on speech using the PZM6FS microphone. Three types of processing are compared: auditory-based pre-processing using delay-and-sum beamforming and the cross-correlation-based enhancement (boxes), auditory-based pre-processing using delay-and-sum beamforming alone (triangles), and LPC processing using delay-and-sum beamforming alone.

6. SUMMARY

The new multi-channel cross-correlation-based processing algorithm was found to preserve vowel spectra in the presence of additive noise and to provide greater recognition accuracy for the SPHINX-I speech recognition system compared to comparable processing of single-channel signals, and compared to comparable processing using delay-and-sum beamforming in the cases examined. We expect to observe further increases in recognition accuracy as further design refinements are introduced to the algorithm.

ACKNOWLEDGMENTS

This research was sponsored by the Defense Advanced Research Projects Agency and monitored by the Space and Naval Warfare Systems Command under Contract N00039-91-C-0158, ARPA Order No. 7239, by NSF Grant IBN 90-22080, and by the Motorola Corporation, which has supported Thomas Sullivan's graduate research. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. We thank Robert Brennan and his colleagues at Applied Speech Technologies for consultations on their multi-channel sampling hardware and software. We also thank the CMU speech group in general and Yoshiaki Ohshima in particular for many helpful conversations, good ideas, and software packages.

REFERENCES

1. Acero, A. and Stern, R. M., "Environmental Robustness in Automatic Speech Recognition", *ICASSP-90*, April 1990, pp. 849-852.

2. Seneff, S., "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing", *Journal of Phonetics*, Vol. 16, No. 1, January 1988, pp. 55-76.

3. Lyon, R. F., "A Computational Model of Filtering, Detection, and Compression in the Cochlea", ICASSP-82, pp. 1282-1285, 1982.

4. Ghitza, O.,"Auditory Nerve Representation as a Front-End for Speech Recognition in a Noisy Environment", *Comp. Speech and Lang*, **1**, pp. 109-130, 1986.

5. Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M., "Complex Sounds and Auditory Images", *Auditory Physiology and Perception*, Cazals, Y., Horner, K., and Demany, L.", Eds., pp. 429-446, Pergamon Press, 1991.

6. Duda, R.O.; Lyon, R.F.; Slaney, M., Correlograms and the Separation of Sounds, *Proc. Twenty-Fourth Asilomar Conference on Signals, Systems and Computers*, **1**, pp. 457-46, Maple Press, 1990.

7. Blauert, J., "Binaural Localization: Multiple Images and Applications in Room- and Electroacoustics", *Localization of Sound: Theory and Applications*, R. W. Gatehouse, Ed., pp. 65-84, Amphora Press, Groton CT, 1982.

8. Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G.W., "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms", *JASA*, Vol. 78, Nov. 1985, pp. 1508-1518.

9. Widrow, B., and Stearns, S. D., *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

10. Peterson, P. M., "Adaptive Array Processing for Multiple Microphone Hearing Aids". RLE TR No. 541, Res. Lab. of Electronics, MIT, Cambridge, MA.

11. Van Compernolle, D., "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings", *ICASSP-90*, April 1990, pp. 833-836.

12. Lyon, R. F., "A Computational Model of Binaural Localization and Separation", *ICASSP-83*, pp. 1148-1151.

13. Stern, R. M., and Trahiotis, C., "Models of Binaural Interaction", *Handbook of Perception and Cognition, Volume 6: Hearing*, B. C. J. Moore, Ed., Academic Press, 1995.

14. Jeffress, L. A., "A Place Theory of Sound Localization", *J. Comp. Physiol. Psychol.*, Vol. 41, 1948, pp. 35-39.

15. Colburn, H. S., "Theory of Binaural Interaction Based on Auditory-Nerve Data. I. General Strategy and Preliminary Results on Interaural Discrimination", *J. Acoust. Soc. Amer.*, **54**, pp. 1458-1470", 1973.

16. Stern, R. M., Jr., and Colburn, H. S., "Theory of Binaural Interaction Based on Auditory-Nerve Data. IV. A Model for Subjective Lateral Position", *J. Acoust. Soc. Amer.*, Vol. 64, 1978, pp. 127-140.

17. Lee, K.F., *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.

Time-Interval Patterns and Auditory Images

Roy D. Patterson and Michael A. Akeroyd

MRC Applied Psychology Unit
15 Chaucer Road, Cambridge, CB2 2EF, U.K.

1 Introduction

The firing of auditory nerve fibers is phase locked to the motion of the basilar membrane at frequencies up to 4-5 kHz (Galambos and Davis, 1943). But relatively little is known about the role of this fine-grain timing information in perception. Most perceptual models begin with a cochlea simulation that generates the fine structure and then promptly removes it in the course of constructing a spectrographic representation of the sound (e.g. Giguerre and Woodland, 1994). It is also the case that little progress has been made in understanding the role of phase in auditory perception. Indeed, it is still commonly assumed that Helmholtz was essentially correct, that phase is of little importance, and that sound quality, or timbre, is largely determined by the distribution of energy across the spectrum (e.g. Moore, 1989, p. 230).

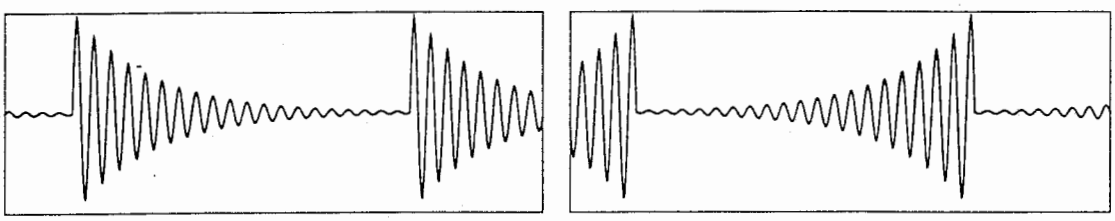
In this paper we argue that the lack of progress in understanding the role of phase locking in perception and the lack of an adequate model of sound quality are related; both arise from the continued implicit use of the Fourier transform as a model of auditory processing. It is argued that the auditory system converts the temporal information in the auditory nerve into some form of multi-channel post-stimulus-time (PST) histogram, and that this form of the 'phase information' provides insight into sound quality that the Fourier phase spectrum does not. The discussion focuses on two new timbre discriminations that are difficult to explain in spectral terms (Section 2). An Auditory Image Model (AIM) of hearing is introduced to convert the sounds into dynamic, multi-channel, PST histograms (Section 3) by means of a simple cochlea simulation and a form of strobed temporal integration. The histograms reveal that there is a time-interval basis for the timbre discriminations, and that the simulated auditory images provide a basis for understanding the sound qualities associated with these stimuli.

2. Timbre Experiments with Tonal Sounds and Noisy Sounds.

2.1 Damped and Ramped Sinusoids

The first timbre contrast involves the sound quality associated with a sinusoid and the effect of asymmetric amplitude modulation on the strength of that quality. The 'damped' sinusoid shown in Figure 1a was produced by applying an exponential decay to a short segment of a sinusoid and then repeating the segment to produce a stationary sound. The frequency of the sinusoid is 800 Hz, the period of the segment is 25 ms, and the half life of the exponential damping function is 4 ms. The 'ramped' sinusoid in Figure 1b was produced simply by reversing the damped sinusoid in time. The damped sinusoid is heard as a unitary source, something like a drum roll on a hollow, resonant object. The ramped sinusoid is heard as a co-ordinated pair of sounds, one of which is like a roll on a non-resonant surface, and the other of which

is *a continuous sinusoid.* The drum role components of the perceptions derive from the streams of transients in these sounds. It is not clear, however, why the ramped sound should carry the distinctive character of the sinusoid and the damped sound not. Patterson (1993) demonstrated the discriminability of damped and ramped sinusoids by presenting listeners with pairs that had the same half life and asking them to choose the member of the pair with the stronger sinusoidal sound quality. The listeners are able to perform the discrimination without difficulty for half lives ranging from 2-16 ms, for carrier frequencies ranging from 400 to 4800 Hz, and for envelope periods ranging from 10 to 100 ms.



(a) (b)

Figure 1. *Waveforms of (a) damped and (b) ramped sinusoids with 4-ms half lives and 800-Hz carriers.*

Auditory models designed to simulate cochlear processing can be used to simulate internal, or auditory, spectra of complex sounds by measuring the level of activity in each channel at a given point in time and plotting the values as a function of channel frequency. Provided the measure is *not* the energy of the filtered wave, the auditory spectra of damped and ramped sinusoids will differ, and if the system is compressive, the peak in the auditory spectrum of the damped sinusoid will be narrower than the peak in the spectrum of the ramped sinusoid. The difference arises because of the way the damped and ramped sinusoids drive off-frequency filters. The damped sinusoid hits the filter with a large pulse of energy which initially causes a strong response. At the same time, however, it causes the filter to try to ring at its centre frequency. Over the course of a few cycles, the ringing energy builds up and, at the same time, the level of the input sinusoid decreases. Since the two terms are associated with different frequencies (the carrier frequency and the filter centre frequency), they eventually drift out of phase and this causes partial cancellation of the output. The ramped sinusoid puts much less energy into the filter initially and when this ringing energy returns to oppose the forcing function, the level of the input sinusoid has increased. The ringing term is small relative to the input throughout the rising portion of the ramped sinusoid. Thus, in a wide range of off-frequency channels, the ramped sinusoid generates more activity at the output when measured in terms of, for example, the average peak-to-trough level, and this difference is preserved in the output of the cochlea simulation. Thus, a model of sound quality based on auditory spectra of this type predicts that damped and ramped sinusoids are discriminable, and that the *damped* member of the pair will sound more like a sinusoid because its spectral peak is narrower and more like that of an unmodulated sinusoid. Nevertheless, the listeners consistently chose the *ramped* sinusoid as the one with the stronger sinusoidal quality.

2.2 Wideband noise versus Iterated Rippled Noise

The second timbre contrast involves the sound of noise. It is the contrast between the sound of a bandpass random noise and that of a bandpass Iterated Rippled Noise (IRN). Rippled noise is constructed from a random noise by delaying a copy of the random noise and adding it back to the original. The delay-and-add process introduces ripples into the spectrum of the IRN, with peaks at multiples of the

reciprocal of the delay and valleys midway between them. The rippled noise sounds like a pair of concurrent sources, a weak low-pitched complex tone in a prominent broadband noise. When the delay-and-add process is repeated, or iterated, the tonal component of the perception grows stronger and the noise component grows weaker, and by about 10 iterations the noise component of the perception is barely noticeable. When the delay is long (say 16 ms), the spectral peaks are closely packed (every 64 Hz) and, in the region above about ten times the peak spacing, the auditory spectrum of the IRN is quite similar to that of the random noise. Thus, a spectral model of hearing would suggest that if random and iterated noises are highpass filtered and equated for energy, they will not be discriminable. Nevertheless, they are perfectly discriminable; one has the *shshsh* of noise and the other sounds like a buzzy musical note. Their phase spectra are complex, random functions that do not provide any obvious explanation for the sound qualities we hear.

The timbre contrast between random and iterated noises led us to suspect that IRN would be more detectable in random noise than in IRN, and that random noise would be more detectable in iterated noise than in random noise. To test this hypothesis, a standard, two-interval forced-choice experiment was performed to determine the signal level required to detect both a random noise and a IRN with 256 iterations in each of five maskers, a random noise and IRNs with 1, 4, 16 and 256 iterations. The IRN was produced using the method described by Yost *et. al.* (1993). This has the effect of making it considerably more difficult to detect the random noise in random noise and the IRN in IRN, because the only cue available to the listeners is the loudness of the sounds. In conditions where there is a timbre contrast, the listeners can ignore the roving level and use the relative level of the tonal and noisy components of the perception. All of the signals and maskers were bandpass filtered between 800 and 4500 Hz and so their long-term auditory spectra all had the same shape. In a spectral model of masking, then, the two signals should be about equally detectable in all of the maskers.

Three normal hearing listeners took part in the experiment and the pattern of results was similar for all three. The average data showed that the random noise was about 4 dB more difficult to detect than the IRN with 256 iterations when the masker was a random noise or an IRN with one iteration. As the number of iterations in the masker increased, however, the IRN signal became about 6 dB harder to detect while the random noise became about 14 dB easier to detect! Thus, the timbre contrasts lead to a 20-dB interaction in masking threshold where few, if any, differences should exist at all.

In summary, the data show that timbre discriminations occur where power spectrum models predict no discrimination. Moreover, the phase spectra of the stimuli do not provide any obvious explanations for the sound quality differences that we hear in these sounds.

3. The Auditory Image Model and Sound Quality

A computational model of peripheral auditory processing has been developed to explain the auditory images we hear when presented with sounds -- the Auditory Image Model of Patterson *et. al.*, (1992a). The image is constructed in three stages, each of which involves filtering, or sorting, the components of the sound in some way, and arranging the products of the process along a space-like dimension. The first two stages simulate the frequency analysis and the laterality analysis performed by the peripheral auditory system in the usual way. Together they produce a frequency-laterality plane like that shown in the centre of Figure 2; in this particular case, it illustrates the separation of a source 40 degrees to the right with energy in the mid-frequencies, from a source 20 degrees to the left with energy at higher and lower frequencies. The frequency-laterality analysis is often presented as if it represented peripheral processing in its entirety. At this point, however, it would be difficult to tell whether the sound was a low-pitched bassoon note or a

bandpass filtered noise with a similar spectral shape, because information about regularity in the fine-structure of the sound is not available in the frequency-laterality plane. The complex sound qualities we hear in tonal sounds indicates that there is a third stage of analysis involving the temporal fine-structure of the neural activity flowing from the cochlea. It is as if the system maintained a two-dimensional array of dynamic PST histograms behind the frequency-laterality plane, one histogram for each frequency-laterality combination. The set of histograms activated by a wideband point source would form a vertical plane at a given angle, like the planes shown in Figure 2. When the sound is periodic, the plane contains a set of regular and related histograms; when the sound is irregular, the plane contains irregular and unrelated histograms. In the Auditory Image Model (AIM), the space described in Figure 2 is the basic space of auditory perception. The images that form in this space are the first internal representation of the sound that we are aware of, and subsequent processing is based on these auditory images.

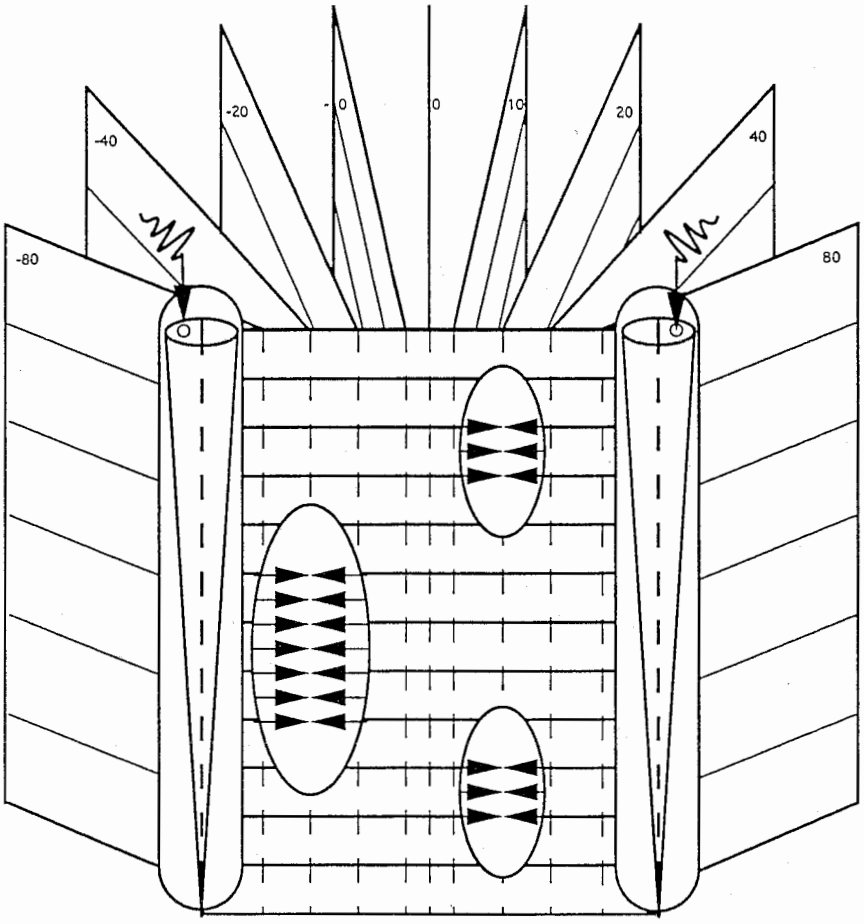


Figure 2. *The space of auditory perception in the auditory image model.*

The mechanism that constructs the histogram from phase-locked neural activity is a new form of temporal integration that is intended to stabilise repeating time-interval patterns from quasi-periodic sounds without smearing their fine-structure. Briefly, a bank of delay lines is used to form a buffer store for the neural activity flowing from the cochlea; the activity level decays as it flows down the buffer at the rate of 2.0 %/ms. Each channel has a strobe unit which monitors the instantaneous activity level and when it encounters a large peak it transfers the entire record in that channel of the buffer to the corresponding channel of a static image buffer, where the record is added, point for point, with whatever is already in that channel of the image buffer. Information in the image buffer decays

exponentially with a half-life of about 30 ms. In the case of periodic and quasi-periodic sounds, the strobe unit tends to synchronise to the period of the sound and so generates a regular stream of pulses that initiate temporal integration in synchrony with the repeating neural pattern. This process stabilises repeating patterns in the simulated neural activity in much the same way as a PST histogram reveals a recurring neural pattern. The one innovation in AIM is that the pulses that reset time in the construction of the PST histogram are derived from the activity pattern itself, without prior knowledge of the sound. When the sound is periodic, or quasi-periodic, this process performs temporal integration over cycles without smearing the fine-structure within the period of the pattern. The traditional leaky integration process removes the majority of the non-energy information from the output of the cochlea simulation. It is argued that the stabilised patterns provide a better basis for analysing sound quality and identifying sources than do spectra, spectrograms or cochleograms (Patterson, *et. al.*, 1992b).

3.1 Auditory Images of Damped and Ramped Sinusoids

The auditory images of the damped and ramped sinusoids with 4-ms half lives and 800-Hz carriers are presented in Figures 3a and 3b, respectively. The figures show the frequency region from 1.5 octaves below, to 1.5 octaves above, 800 Hz (9.3-19.2 ERBs). In the upper section of each sub-figure, the activity in each channel is essentially an impulse response produced by the abrupt change in amplitude of the stimulus once per cycle of the sound. The response to the carrier frequency is shown in the central section of each panel. The ramped sinusoid (Figure 3b) activates a broader frequency region than the damped sinusoid (Figure 3a). The phase alignment induced by strobed temporal integration reveals that the carrier activity of the ramped sinusoid is mainly composed of time intervals at the period of the carrier, even when the activity is in channels well above or below the carrier channel. This sound has the quality of a sinusoid. The response to the damped sinusoid reveals time intervals characteristic of the carrier only in the channel at the centre of the carrier response. As the channel frequency decreases below that of the carrier, the time intervals in the response lengthen, and as the channel frequency increases above that of the carrier, the time intervals shorten. This behaviour is characteristic of a set of resonators struck by an acoustic pulse; each rings at its own centre frequency. This sound has a hollow quality rather than a sinusoidal quality.

Damped and ramped sounds produce relatively simple auditory images composed either of time intervals at the carrier period or at the period of the centre frequency of the channel. For these simple images, it is possible to segregate and measure the carrier period activity, and so test the hypothesis that the sound of a sinusoid is associated with a concentration of time intervals at the carrier period rather than with a concentration of energy in the carrier channel. Auditory images were produced for a range of damped and ramped sounds and time-interval histograms were calculated for each channel of each image. (The calculations were limited to time-intervals between adjacent pulses for simplicity.) The multi-channel time-interval histograms confirmed that the ramped sinusoid produces a substantial number of 1.25-ms intervals in off-frequency channels whereas the damped sinusoid does not. The number of 1.25-ms intervals was calculated for each auditory image and pairs of values for damped and ramped sinusoids with the same half life were compared.

Broadly speaking, for all carrier frequencies, the number of carrier periods was substantially higher for the ramped sinusoid in the range of half lives where discrimination performance was good, that is, between 2 and 16 ms. For the higher carrier frequencies (3200 and 4800 Hz), the model also produced more carrier periods for the ramped sinusoid at half lives less than 2 ms, where the listeners could not hear a difference. This presumably reflects a loss of phase locking at short time-intervals in the auditory system; a loss which is not simulated in the current version of AIM. Despite its obvious limitations, the quantitative analysis of carrier period activity suggests that it is the presence of time intervals at the carrier period

rather than energy in the carrier channels *per se* that gives a sound the quality associated with a sinusoid.

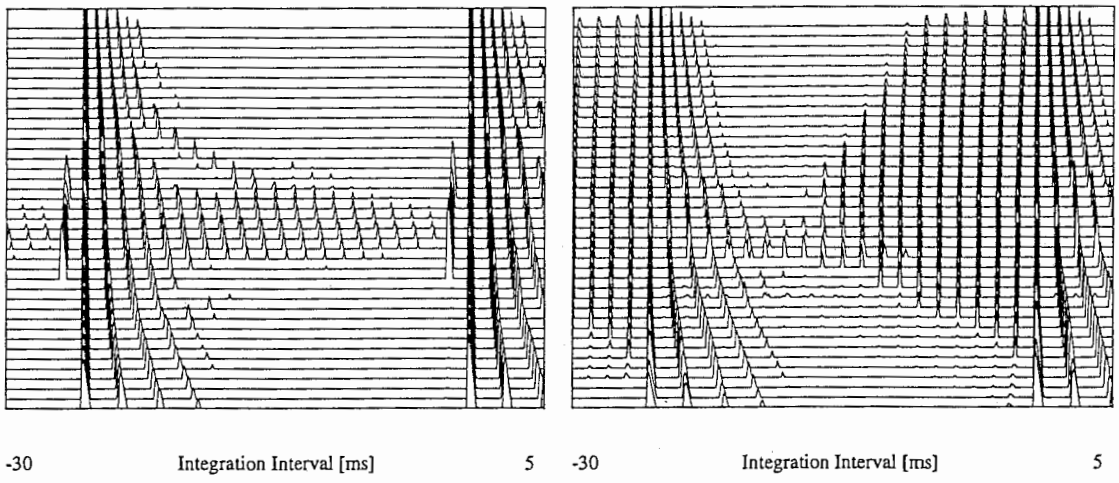


Figure 3. *Auditory images of (a) damped and (b) ramped sinusoids with 4-ms half lives. The time interval between the parallel vertical ridges in ramped image is 1.25 ms -- the period of the carrier.*

3.2 Auditory Images of Wideband noise and Iterated Rippled Noise

The second timbre contrast was between the sound of a bandpass random noise and the sound of a bandpass IRN. Auditory images of a white noise and a ripple noise with 256 iterations and a 16-ms delay are presented in Figure 4a and 4b. The frequency region is once again 9.3-19.2 ERBs. Both images show regularity at the shortest time intervals, since even white noise is correlated with itself in the short term. Beyond a few cycles, however, there are no stable features in the white noise image (Figure 4a). In the IRN, there are vertical ridges at 16 and 32 ms (Figure 4b) indicating temporal regularity both within and across channels at these time intervals. The auditory images provide a basis for understanding the

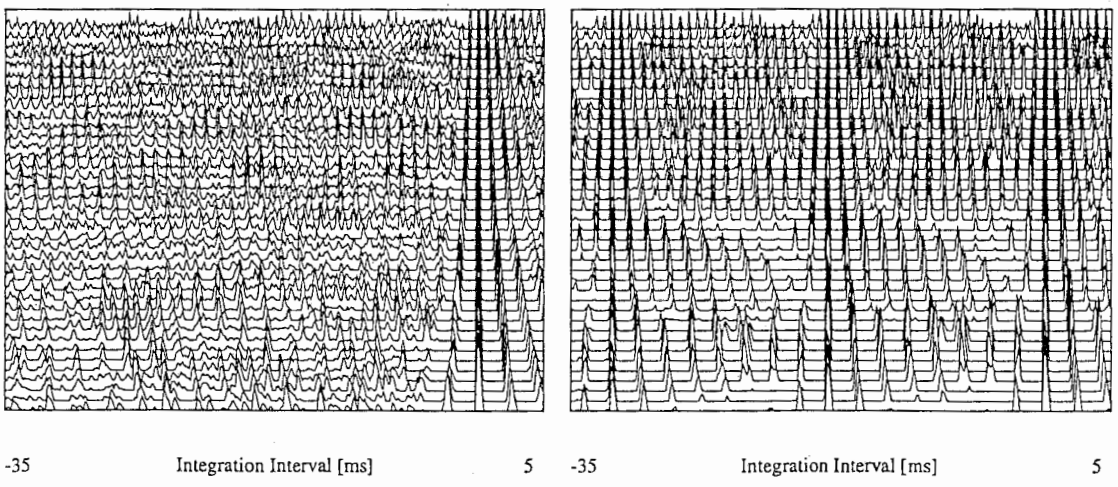


Figure 4. *Simulated auditory images of (a) random noise and (b) IRN with 256 iterations.*

timbre of these sounds. The *shshsh* of noise arises when the image contains regions where there is no structure and the rate of change in the detail is relatively high. In

contrast, a tonal sensation arises when the image contains regions where the time-interval pattern is orderly and the rate of change in the detail is relatively low. If the auditory system performs an analysis of time-interval patterns like that implied by these simulated auditory images, and if it is able to analyse the images and separate the orderly regions (i.e. figures) from the unordered regions(i.e. background noise), then the ratio of the level of activity in the figure to the level of activity in the background would provide a basis for the timbre discrimination data from the IRN experiment. When the signal and masker have the same number of iterations adding the signal to the noise does not alter the form of the auditory image; there is nothing but a level difference to distinguish the interval with the signal, and the roving level paradigm makes this difficult. But when the signal has no iterations and the masker has many, the signal can be detected by the irregularity that it introduces into the auditory image associated with the signal interval (Figure 4b). The roving level paradigm does not interfere with detection in this case because the listener is using the figure/ground ratio as the cue to the presence of the signal. We do not yet have an algorithm for identifying and segregating regular and irregular regions of auditory images, and so we do not have a quantitative measure of the figure/ground ratio to compare with the data from the experiment. Nevertheless, this qualitative explanation of the timbre discrimination would appear to be better than what could be expected from a spectrographic model of sound quality for these stimuli.

4 Summary

The auditory image model has been used to convert the phase-locked time-interval patterns of some tonal and noisy sounds into dynamic PST histograms, and to relate the patterns in the histograms to timbre discriminations. The first discrimination involved the strength of the sinusoidal quality produced by asymmetrically modulated sinusoids, which was found to vary with the direction of the asymmetry. The model explains that the onsets of sinusoids drive off-frequency auditory filters in a temporally synchronous mode whereas the offsets of sinusoids do not. The auditory images of damped and ramped sinusoids suggest that it is these time-intervals at the carrier period that give rise to the sound of a sinusoid rather than energy in the carrier channel *per se*. The second discrimination involved the difference between the *shshsh* of random noise and the tonal buzz of iterated rippled noise. It led to a masking experiment where the timbre contrast was found to support a 20-dB interaction in masking levels that would not have been anticipated from the spectra of the sounds. The discrimination was explained in terms of the variability of the time intervals in the multi-channel histograms of the sounds. Together the model and experiments suggest that the timbre of sounds is closely related to the time-interval patterns produced by sounds in the auditory nerve, and that even simple models of the production and processing of these time-interval patterns are sufficient to support productive research into the perception of sound quality.

4.1 Acknowledgements

The authors would like to thank W. Yost for assistance with the IRN generators; J. Datta and S. Handel for contributions to the IRN experiment and comments on an earlier draft of the paper; M. Allerhand for continuing support of the AIM software; and M. d'Souza for help with the postscript figures. The research was supported by DRA grant FRNIC/U/759 and an MRC studentship to author M. A.

4.2 References

Akeroyd, M.A. and Patterson, R.D. (1994) Discrimination of time-asymmetric modulated noise. J. Acoust. Soc. Am. MIT meeting, (in press).

Giguerre, C. and Woodland, P.C. (1994) A computational model of the auditory periphery for speech and hearing research: I. Ascending path. J. Acoust. Soc. Am. 331-342.

Galambos, R. and Davis, H. (1943) The response of single auditory nerve fibers to acoustic stimulation. J. Neurophysiology, 6, 39-57.

Moore, B.C.J. (1989) An introduction to the psychology of hearing. Academic Press, London.

Patterson, R. (1993) What determines the sound of a sinusoid? J. Acoust. Soc. Am. 93, 2293 (A) (1993).

Patterson, R.D. (1987) A pulse ribbon model of monaural phase perception. J. Acoust. Soc. Am. 82, 1560-1586.

Patterson, R.D., Holdsworth, J. and Allerhand M. (1992b) 'Auditory Models as preprocessors for speech recognition', In: *The Auditory Processing of Speech: From the auditory periphery to words,* M. E. H. Schouten (Ed), Mouton de Gruyter, Berlin, 67-83.

Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand, M. (1992a) Complex sounds and auditory images. In: Auditory physiology and perception. Eds. Y Cazals, L. Demany, K. Horner, (Pergamon, Oxford), 429-446.

Yost, W.A., Allerhand, M., Robinson, K., and Patterson, R.D. (1993) The pitch and pitch strength of iterated rippled noise. Abstract of the 16th meeting of the Association. for Research in Otolaryngology, No. 186.

A TEMPORAL ACCOUNT OF COMPLEX PITCH

William A. Yost, Stanley Sheft, Bill Shofner
Parmly Hearing Institute, Loyola University of Chicago, Chicago IL

and Roy Patterson
Applied Psychology Unit, MRC, Cambridge England.

In recent years several hearing scientists have described the crucial role that sound source determination or sound source segregation plays in hearing (see for example Bregman, 1990; Hartmann, 1988; or Yost, 1992). An important variable in allowing the auditory system to determine the source of sound is the auditory system's sensitivity to the harmonic structure of many sounds in our everyday world. The "case of the missing fundamental pitch" spawn the realization that many complex stimuli produce a pitch that is not a simple transform of the spectral or temporal characteristics of the waveform. These stimuli characterize many sounds that occur in our everyday lives, including speech. The pitches produced by these complex sounds have been labeled complex or virtual pitch. For many, if not most, of these stimuli the complex pitch of the sound occurs along with other perceptual attributes. That is, besides the complex pitch the sound may also have a "tinny" or "noisy" timbre. Sometimes it is as if there are two potential sound sources: that producing the complex pitch and a second source that is responsible for the other timbrel percept. Since a major role of hearing is the segregation of the various sources that make up a complex sound scene, these complex sounds offer a potential advantage for studying sound source segregation.

In our presentation, we describe a class of complex pitch stimuli which we call iterated ripple noise (IRN). IRN produces a complex pitch sometimes called "repetition pitch" (see Bilsen and Ritsma, 1970; or Yost and Hill, 1978), but in addition the IRN stimuli have a noisy timbre that appears along with the repetition pitch much as if there were two sound sources that generated the IRN stimulus. IRN typifies most complex pitch stimuli. We present human psychophysical evidence that IRN is processed temporally and not spectrally. In addition we present some physiological data from chinchillas showing the neural temporal sensitivity of units in the cochlear nucleus to IRN. We show that the chinchilla is also psychophysically processing ripple noise similarly to the way humans do. And, finally we show that autocorrelation of the stimulus, which can be form by an auditory correlogram or by the auditory image model (AIM) of Patterson and Holdsworth (Patterson et al, 1992), can account for essentially all of the data.

Living in a reverberant world, we constantly experience the sound from its source plus its many echoes. The perception of a sound and an echo is a "spectral coloration" or a "repetition pitch" added to the sound of the input. When a noise is introduced to an add and delay network, two stimuli are generated which have been used to study spectral coloration: cosine noise (see Yost et al, 1978) and comb noise (see Raatgever and Bilsen, 1983). In these two networks the original noise is delayed and added back to itself (after undergoing some attenuation), much as would occur for a sound and its echo. The Fourier transform [H(w)] of the feedback network used to generate comb noise is:

$$H(w) = 1 + g\exp(-jwT) + g^2\exp(-j2wT) + \ldots + g^{n-1}\exp(-j(n-1)wT); \qquad 1)$$

where g ($0 \leq g \leq 1$) is attenuation, $w=2f\pi$, T is the delay, n = number of iterations

and for cosine noise the network Fourier transform is:

$$H(w) = 1 + g \exp(-jwT). \qquad 2)$$

As can be seen from equations 1) and 2), comb noise is infinitely iterating the add and delay network used to generate cosine noise. Here the output of each delay and attenuate circuit is added back to the original input (this network will be referred to as the add-original network).

In a different network used to generate IRN, the output of each delay and attenuate is added back to the previous added waveform, and not to the original input (this network will be referred to as the add-same network). The Fourier transform of the add-same network is:

$$H(w) = [1 + g \exp(-jwT)]^n; \qquad 3)$$

Figure 1 shows these networks.

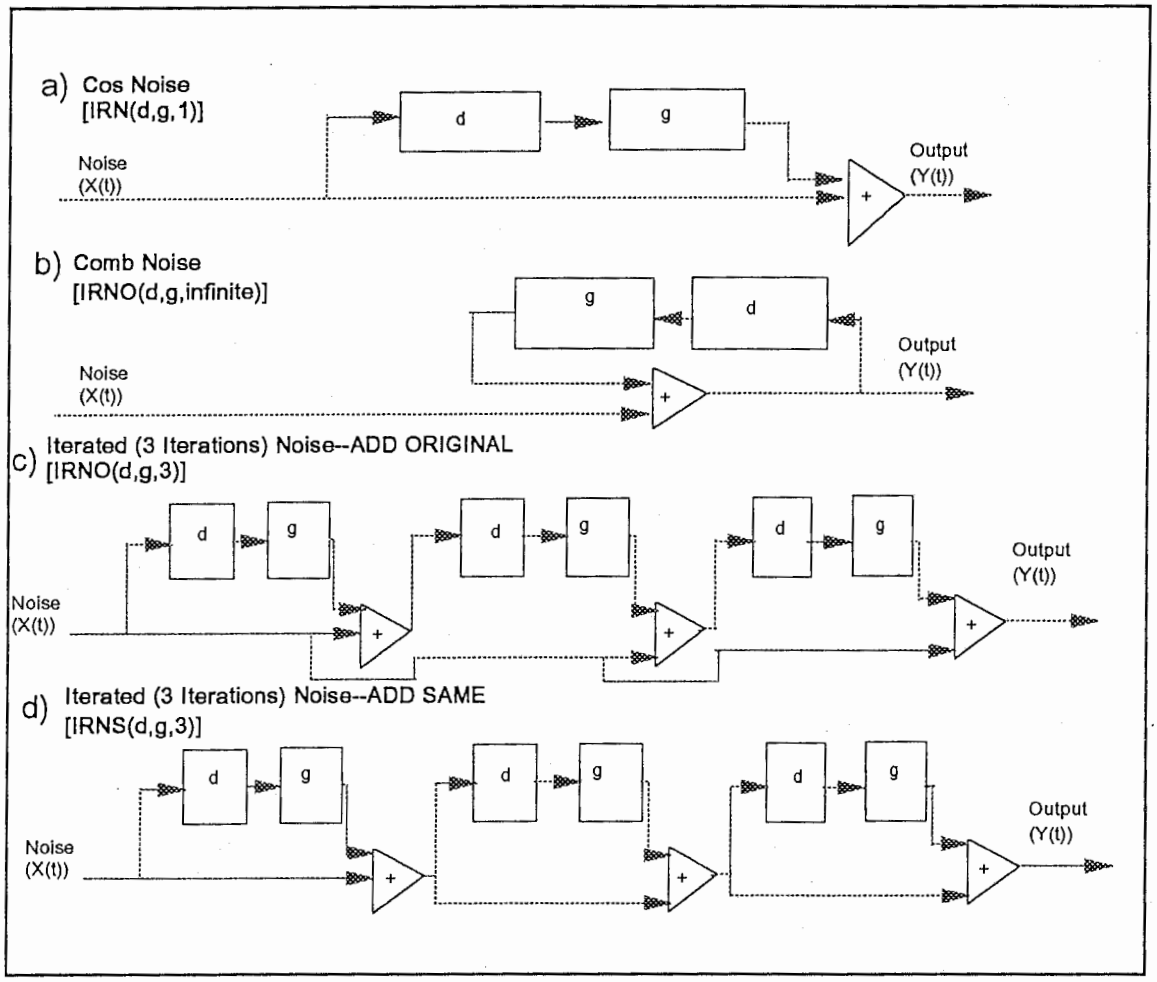


Figure 1. The add, delay (d), and attenuate (g) networks used to generate: a) Cosine Noise (IRN(d,g,1)), b) Comb Noise (IRNO(d,g,∞)), c) Iterated Noise in the ADD ORIGINAL network (IRNO(d,g,3)) with three iterations, and d) Iterated Noise in the ADD SAME network (IRNS(d,g,3)) with three iterations.

There is spectral ripple of power as a function of frequency, such that the major peaks in the spectra are at integer multiplies of 1/T. In the add-original network the spectral peaks sharpen and the number of smaller spectral peaks in the valleys between the major peaks increases and their amplitudes decreases as the number of iterations increase. In the add-same network there are no smaller spectral peaks in the valleys between the major peaks and the major spectral peaks also sharpen as the number of iterations increases. The peaks in the spectra decrease and the amplitudes in the valleys increase as g is lowered toward zero (i.e., when g=0 the spectra are flat like that of the original noise input).

As the number of iterations increases in each network, the strength of the repetition pitch of IRN varies along a continuum from a noisy sound with a subtle pitch to a sound with a pitch like that of a pulse train but without a noisy timbre. It is as if each network produces two sound sources, a noise and a complex tone. The noise source becomes less salient and the tonal complex more salient as the number of iterations increases.

Cosine noise has been used to study many aspects of complex pitch (see Yost et al, 1978), to derive critical band functions (Houtgast, 1973), to study after images (Wilson, 1969), and to investigate the perceptions of a sound and its echo (Bassett and Eastmond, 1964). By varying the delay, T, used to generate cosine noise, its virtual pitch equals 1/T. By varying the value of g, the strength of the pitch of cosine noise varies from a hollow pitch in a noisy background when g=1 to a noise without a pitch when g is near zero. Studies of cosine noise have provided a means of studying both complex pitch and complex pitch strength, and these studies (see Yost et al, 1978) have documented that cosine noise is one of the classes of stimuli that generate complex or virtual pitch. Since cosine noise has no periodicities in its time waveform and it has a continuous spectrum, it has proven a difficult stimulus for many models of complex pitch (see Yost, 1979).

Far fewer studies of hearing have used comb noise (Bilsen and Weiman, 1980; Fastl, 1988; and Raatgever and Bilsen, 1983), but it also produces a virtual pitch equal to 1/T and its pitch strength diminishes as g approaches zero (see Raatgever and Bilsen, 1992). However, the pitch strength of comb noise is considerably stronger than that of cosine noise when g is near 1.0. The perception of comb noise is much like that of a pulse train with a repetition period of T. Only a few studies have investigated iterated ripple noise (Yost et al, 1993), and it too has a pitch equal to 1/T independent of the number of iterations (Yost et al , 1993), and as explained above its pitch strength increases as the number of iterations increases and as g approaches 1.0.

We will report on a series of experiments investigating the discrimination between various IRNs as a function of number of iterations, attenuation, and type of network. In general, this is a study of the pitch strength or saliency of IRN. The research will show that the pitch strength of IRN is most likely dependent on the temporal properties of the waveform. When listeners were asked to discriminate between certain pairs of IRN, some stimuli were almost impossible to discriminate from each other despite large spectral differences, while for other comparisons discrimination was easy despite small spectral differences. No simple spectral processing rule can account for the results. A simple temporal transform, based on autocorrelation, will be shown to be consistent with the results of all experiments. These results strongly suggest that the underlying processing of these complex pitch stimuli is temporal and not spectral, and that sometimes the auditory system cannot process spectral differences even when they are substantial.

For both types of network, peaks in the autocorrelation functions appear at lags of mT (m=1,2,3...,n; n is the number of iterations), and the magnitude of the peaks decreases for increasing

m. The heights of the first peak (lag=T) in the normalized autocorrelation functions are the same in the two networks for the same number of iterations, but the height of the peaks at integer multiplies of T are smaller in the add-same than in the add-original networks. The heights of the autocorrelation peaks in both networks increase with increasing n and decrease as g approaches zero.

The general finding from all of the human psychophysical work is that discrimination performance and the perceived magnitude of pitch strength can be almost perfectly predicted from the heights of the first peak in the autocorrelation functions. The height of the first peak in the normalized autocorrelation function indicates the relative proportion of intervals of length T in the fine structure of the waveform with all other intervals being randomly distributed. These intervals of constant duration T are abundant in IRN, but they do not occur periodically. The number of intervals increases as the number or iterations increases, but decreases for any one IRN when g is decreased toward zero.

When two waveforms have the same number of autocorrelation peaks and the heights of the first peak in the autocorrelation functions are equal (the case of comparing an IRN waveform in the add-same with one in the add-original network each produced with the same n, T, and g), the two sounds are indiscriminable from each other although there are large spectral differences and the heights of the peaks in the autocorrelation function after the first peak are different. The spectra associated with these comparisons and the discrimination results are shown in Fig. 2

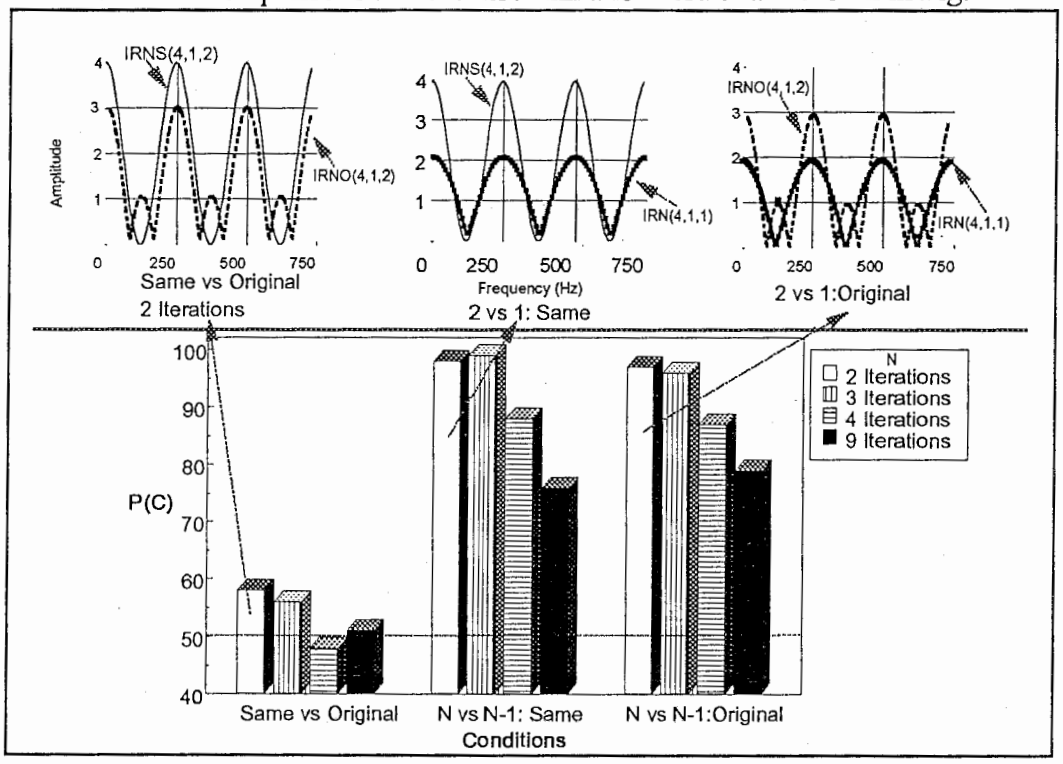


Figure 2. The panels at the top show pairs of amplitude spectra for stimuli to be discriminated in three experimental conditions. These are the amplitude spectra of the network transfer functions. The solid curves are the spectral transforms for IRNS(2,1,2), the dashed curves are for IRNO(2,1,2), and the curves with the squares for IRN(2,1,1). The bottom panel shows the mean P(C) value (over 7 listeners) for discriminating IRNS from IRNO which is difficult, and for discriminating IRNS(d,1,n) from (IRNS(d,1,n-1) and IRNO(d,1,n) from IRNO(d,1,n-1) both of which are easy.

In other cases the heights of the first peaks of two IRNs can be varied by using different values of g and n, although the number of peaks in the autocorrelation functions may vary from stimulus to stimulus (i.e., comparing two IRNs generated in the add-original network each produced with a different n and g). Both discrimination data and pitch strength scaling data are extremely well described by using the height of the first peak in the autocorrelation function as the predictor of performance. The differences and similarities in the autocorrelation functions appear much more consistent with all of the experimental outcomes than any simple spectral comparisons. However, since autocorrelation is the Fourier transform of the power spectrum, there are spectral comparisons that are consistent with the changes in the height of the first peak in the autocorrelation function. However, both the results and some controls introduced into the experiments appear to rule out a simple spectral explanation of the data.

In our work with chinchillas, we show that they can discriminate differences between various pairs of IRNs with essentially the same shaped psychometric functions as those produced by humans, but the chinchillas are far less sensitive to changes in the IRNs than are humans. That is, the chinchilla's psychophysical data can be accounted for on the basis of the first peak in the autocorrelation function, but much larger changes are necessary to account for their data than are needed to account for the human data. Figure 3 shows the results of an experiment in which the chinchillas discriminated between a flat noise and a cosine noise and between a flat noise and a comb filtered noise with the gain in the network set to 0.5. To humans these two iterated rippled noises (the cosine and the comb noise) are difficult to tell apart since the first peak in the autocorrelation functions are nearly equal. It is difficult to ask animals to work in a discrimination task where the stimuli are difficult to discriminate. In the comparison in Fig. 3, the two iterated noises are each equally discriminable from a flat noise, and thus it is reasonable to assume that the two iterated ripple noises would be difficult for the chinchillas to discriminate.

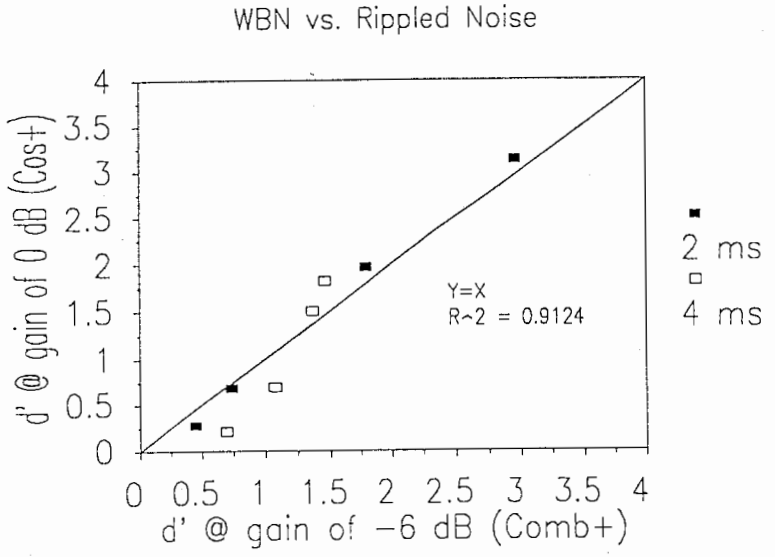


Figure 3. The d' values for four chinchillas in making a discrimination between a flat noise and cosine ripple noise with g = 0, and between a flat noise and comb-filtered ripple noise with g=-6dB. The diagonal line represents equal discrimination for each condition. As can be seen the two discriminations produce about the same discrimination performance for each delay used (2 and 4 ms). Cosine noise and comb-filtered noise with g=-6 dB would be difficult for humans to discriminate.

The renewal density function for many neural units in the cochlear nucleus of the chinchilla parallel closely the autocorrelation functions and as such demonstrate that these neural units preserve the temporal information necessary to account for the human and the chinchilla's psychophysical performance. The data in Fig.4 show the neural responses plotted as neural renewal density functions. A renewal density function is basically an autocorrelation function, where autocorrelation magnitude can be expressed in terms of spikes/sec. The three lines on the graph represent the mean spike rate plus and minus one standard deviation. Thus, any peak in the renewal density function above the upper line is seen as being statistically different from the average firing of the neuron. These peaks occur at integer multiplies of the delay used to generate this comb filtered noise.

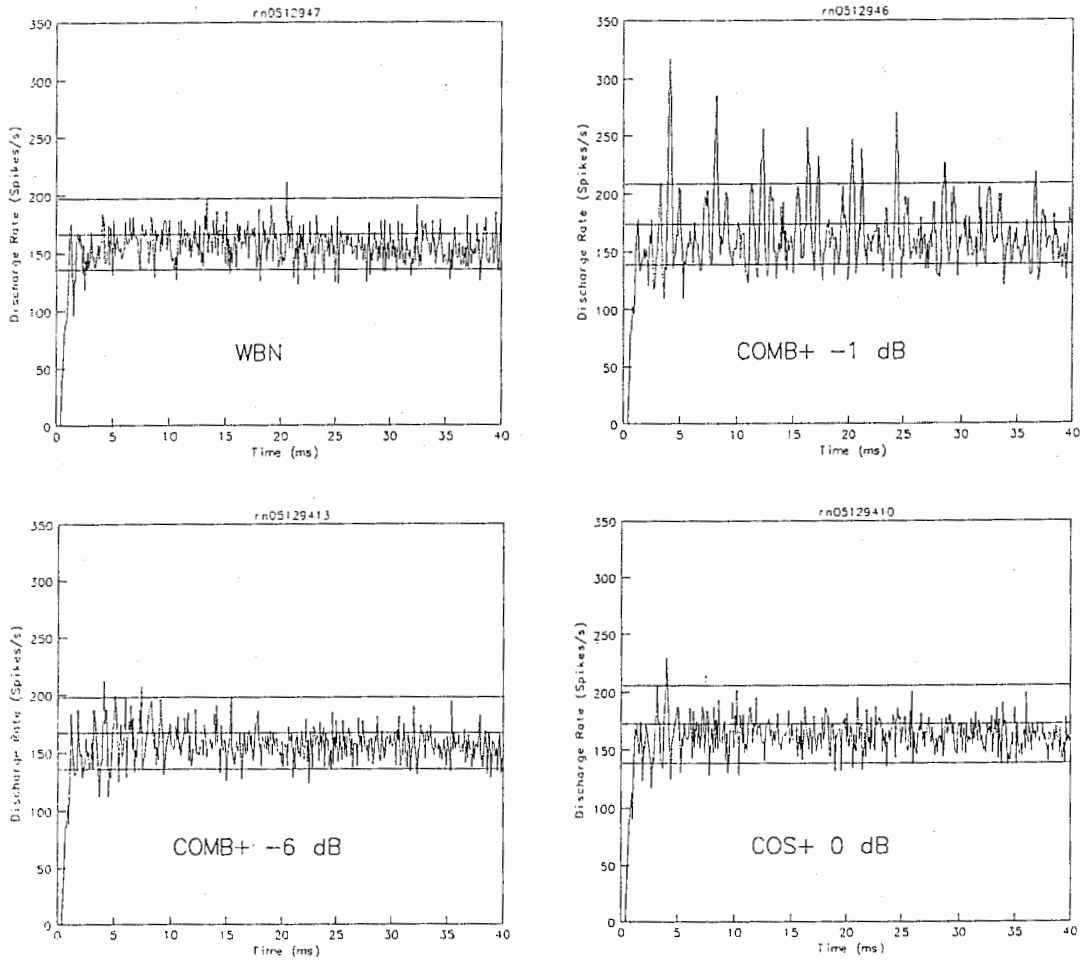


Figure 4. The renewal density function for a primary like neuron in the cochlear nucleus of the chinchilla. The upper left shows the data for wideband noise, the upper right for comb-filtered noise with g=-1 dB, the lower right comb-filtered noise with g=-6dB, and the lower right cosine noise with g=0dB. Any peaks above the upper horizontal line (representing mean firing rate plus one standard deviation) represents a firing rate at that interval that is significantly above the mean firing rate. The peaks are significant at intervals of 4 ms and its integer multiplies, which is the delay used to generate the ripple noises. The lower two functions look similar with only one peak at 4 ms being significant. Humans and chinchillas (see Fig. 3) have difficulty discriminating between these two ripple noises.

The use of autocorrelation has been suggested previously for explaining the pitch and pitch strength of complex pitch (Wightman, 1973; Patterson and Wightman, 1976; and Yost, 1979), especially the repetition pitch of cosine noise (Yost, 1979). In general these models assume that the location of peaks in the autocorrelation function at different lags determines the virtual pitch of the stimulus and the height of the autocorrelation peaks determines the relative pitch strengths. The results of the present experiments are entirely consistent with the assumptions of these models. The current results further show that discriminability among complex pitch stimuli based on pitch strength can be accounted for by the autocorrelation functions. The data also suggest that it is unlikely that a simple spectral explanation can account for the discrimination results. Thus, this research strongly supports a temporal interpretation of the processing of complex pitch.

Designing a neural delay-line circuit to perform autocorrelation has been suggested often since the early work of Jeffress (1948) and Licklider (1956). However, the computations required of such a circuit have to be fairly extensive to maintain the temporal accuracy necessary to account for the data of this and related complex pitch studies. The same is true of the simulations based on weighted autocorrelation (auditory correlogram) used in models such as those proposed by Meddis and colleagues (1991) and by Lyon and Slaney (1991). These models have had success in accounting for the pitch of cosine noise. The stabilized auditory image (SAI) stage of the Patterson and Holdsworth model (Patterson et al, 1992) suggests a simple and efficient trigger method for extracting temporal coincidences such as those occurring in IRN. As such the SAI stage performs a form of autocorrelation which is consistent with the data of our study.

In studies of neural units in the goldfish (Fay et al, 1983), cat (Boerger, 1974), and chinchilla (Shofner, 1991) in response to cosine noise, it has been shown that many neurons in the auditory nerve and cochlear nucleus have neural responses whose interval histograms, autocorrelation functions, and renewal density functions display temporal information like those described above. That is, the neurons discharge at intervals of T or integer multiplies of T. To obtain such temporal properties, the spike-train data from many replications of each stimulus were averaged. That is, the data are averaged over time. The computations based on the SAI model or correlograms also suggest that averaging over different frequency channels can reveal the temporal structure of IRN.

The overall conclusion of this paper is that the pitch strength of IRN can be explained by using the autocorrelation function of the stimulus. The results further demonstrate that for some of these IRNs the auditory system is unable to discriminate one noise from another despite relatively large spectral differences, while for other comparisons discrimination is easy despite small spectral differences. These results support a temporal explanation for processing complex pitch stimuli.

Acknowledgment. This research was supported by grants from the NIDCD, AFOSR, and DRA Farnborough. We would like to thank Drs. Sheryl Coombs, Toby Dye, Bill Shofner, and Steven Handel for their advice regarding this research. Portions of these data were reported on at the 127th Meeting of the Acoustical Society of America and at the Hearing Conference in Irsee Germany in July, 1994.

References

Bassett, I.G. and Eastmond, E.J. (1964) Echolocation: Measurement of pitch versus distance for sounds reflected from a flat surface, J. Acoust. Soc. Am. 36, 911-916.

Bilsen, F. A. and Ritsma, R. J. (1970). Some Parameters Influencing the Perceptibility of Pitch, J. Acous. Soc. Am. 47, 469-476.

Bilsen, F. A. and Wieman, J. G. (1980). Atonal Periodicity Sensation for Comb Filtered Noise Signals, in Psychophysical, Psychological and Behavioral Studies in Hearing, edited by van Den Brink and F. A. Bilsen (Delft University Press,).

Boerger, G. (1974). Coding of Repetition Noise in the Cochlear Nucleus in the Cat, in Facts and Models in Hearing, edited by Zwicker and Terhardt (Springer-Verlag, New York).

Bregman, A.S. (1990) Auditory Scene Analysis. MIT Press, Cambridge, MA.

Fastl, H. (1988) Pitch and Pitch Strength of Peaked Noise. In Basic Issues in Hearing, D. Duifuis, H.P. Wit, and J.W. Horst (eds), Academic Press.

Fay, R.R., Yost, W.A., and Coombs, S. (1983) Psychophysics and Neurophysiology of Repetition Noise Processing in a Vertebrate Auditory System, Hear. Res. 12, 31-56.

Jeffress, L. A. (1948). A Place Theory of Sound Localization. The Journal of Comp. and Physio. Psychol., 35-39.

Hartmann, W. (1988) Pitch Perception and the Organization and Integration of Auditory Entities. In: G.W. Edelman, W.E. Gall and W.M. Cowan (Eds.),Auditory Function: Neurobiological Bases of Hearing. John Wiley and Sons, New York.

Houtgast, T. (1973). Psychophysical Experiments on Tuning Curves and Two-Tone Inhibition, Acustica 29, 168-179.

Licklider, J. C. R. (1956). Auditory Frequency Analysis, in Information Theory, edited by C. Cherry (Butterworth Press, London), 253-268.

Meddis, R. and Hewitt, T. (1991) Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery I: Pitch Identification, J. Acoust. Soc. Am. 89, 1862-1882.

Moore, B.C.J. (1993) Frequency in Psychoacoustics, Yost, W.A., Fay, R.R., Popper, A. (Co-Editors), Springer Verlag.

Patterson, R. D. and Wightman, F. L. (1976). Residue Pitch as a Function of Component Spacing, J. Acoust. Soc. Am.59, 1450-1460.

Patterson RD, Robinson K, Holdsworth J, McKeown D, Zhang C, Allerhand M (1992) Complex Sounds and Auditory Images, Auditory Physiology and Perception (eds. Y. Cazals, K. Horner, and L.Demany) Pergamon Press, Oxford.

Raatgever, J. and Bilsen, F.A. (1992) The Pitch of aAharmonic Comb Filtered Noise Reconsider. Auditory Physiology and Perception (eds. Y. Cazals, K. Horner, and L.Demany),Pergamon Press, Oxford.

Shofner, W.P. (1991), Temporal Representation of Rippled Noise in the Anteroventral Cochlear Nucleus of the Chinchilla, J. Acoust. Soc. Am. 90, 2450-2466.

Slaney, M and Lyon, R.F. (1991) Apple Hearing Demo Reel, Apple Technical Report, Apple Computer, Inc.

Wightman, F. L. (1973). Pattern-Transformation Model of Pitch, J. Acoust. Soc. Am. 54,

407-417.

Wilson, R (1969) An Auditory After-Image, in Frequency Analysis and Periodicity Detection in Hearing, R. Plomp, G.F. Smoorenburg (eds), Sijthoft Press, Leiden.

Yost, W. A.(1992a), Auditory Image Perception and Analysis, Hearing Research 56, 8-19.

Yost, W. A., (1979) Models of the Pitch and Pitch Strength of Ripple Noise, J. Acoust. Soc. Am. 66, 400-411.

Yost, W. A. and Hill, R., Strength of the Pitches Associated with Ripple Noise, J. Acoust. Soc. Am., 64, 485-492, 1978

Yost, W.A., Hill, R., and Perez-Falcon, T. (1978) Pitch and pitch discrimination of broadband signals with rippled power spectra, J. Acoust. Soc. Am. 63, 1166-1173.

Yost, W. A., Allerhand, M., Zhang, Robinson, K., and Patterson, R., (1993) The Pitch and Pitch Strength of Iterated Ripple Noise, Assoc. Res. Oto. Abst.

EXTRACTING THE FUNDAMENTAL FREQUENCIES OF TWO CONCURRENT SOUNDS

Robert P. Carlyon, MRC Applied Psychology Unit, 15 Chaucer Rd., Cambridge CB2 2EF, England

One of the most demanding situations that the auditory system is confronted with occurs when two or more concurrent sounds, such as the voices of two speakers, are presented simultaneously. In order to identify what either voice is saying, listeners must determine which components of the combined spectrum correspond to which voice. When performing this task, listeners exploit the fact that components of any periodic source have frequencies which are harmonics of a common fundamental frequency ("F0": Brokx and Nooteboom, 1982; Scheffers, 1983). The importance of F0 differences ("ΔF0s") for the segregation of concurrent sounds has been further established by a number of studies by speech scientists (Scheffers, 1983; Assmann and Summerfield, 1990) , psychoacousticians (Carlyon *et al.*, 1992; Carlyon and Shackleton, 1994), auditory modellers (Assmann and Summerfield, 1990; Cooke, 1992; Meddis and Hewitt, 1992) , and signal-processing engineers (Parsons, 1976; Stubbs and Summerfield, 1988; Slaney and Lyon, 1990).

The experiments described below used greatly simplified speech-like stimuli to investigate two ways in which the auditory system might use ΔF0s when segregating concurrent sounds. The first set of experiments investigate our ability to compare the F0s of two groups of components (such as formants) which are well-separated in frequency. These experiments, which have important implications for models of F0 encoding, will be summarised only briefly, as they have been described elsewhere (Carlyon and Shackleton, 1994). The second set, which will be described in more detail, measure the extraction of the F0s of two groups of harmonics which are mixed in a single frequency region.

I. DETECTING ΔF0s BETWEEN RESOLVED AND UNRESOLVED HARMONICS IN DIFFERENT FREQUENCY REGIONS

Consider two formants which are well-separated in frequency, where the harmonics of one formant are resolved by the peripheral auditory system, but where those of the other are unresolved. How can the listener tell whether these are formants of the same voice, with a common F0, or whether they have different F0s, and therefore come from different voices? According to traditional theory, the F0s of resolved and unresolved harmonics are processed via separate mechanisms, and so a comparison of the two F0s should not be straightforward. In contrast, more recent theories (Meddis and Hewitt, 1991; Patterson *et al.*, 1991) propose a common mechanism for encoding the F0s of resolved and unresolved harmonics, thereby predicting no special difficulty in this particular task. Carlyon and Shackleton (1994) investigated whether the F0s of resolved and unresolved harmonics are encoded by the same or by different mechanisms. They used a stimulus consisting of two formant-like groups of components, well-separated in frequency, and presented in a noise background in order to mask within-channel interactions. They found that listeners were much worse at detecting ΔF0s between a group of resolved harmonics and a group of unresolved harmonics, compared to the case where the two groups were either both resolved or both unresolved. They concluded that different mechanisms encode the F0s of resolved and of unresolved harmonics, a conclusion consistent both with the traditional view and with a modern computational approach adopted by Cooke (1992).

II. EXTRACTING TWO F0S FROM THE SAME FREQUENCY REGION

In Carlyon and Shackleton's experiments, listeners were required to compare the F0s of two sets of harmonics which occupied discrete frequency regions. Such across-frequency comparisons are useful when there are regions of the combined spectrum

which contain energy from only one source, and where the listener needs to assign those regions to the appropriate sound - for example, where the formants of two vowels from different speakers are interleaved across the frequency spectrum. However, in many situations, formants are not simply interleaved: instead, there are frequency regions which contain energy from both sources. It would therefore seem useful for the listener to be able to extract more than one F0 from the same frequency region.

Initially, we performed an informal pilot experiment to determine whether listeners can indeed perform this task. They were presented with a 500-ms harmonic complex, consisting of a large number of consecutive harmonics of 210 Hz which had been bandpass filtered. After the first 150 ms of this sound, a second harmonic complex. with the same level as the first and passed through an identical filter, was added; its F0 was either 175 or 252 Hz and its duration was 200 ms. The perceptual results of this simple manipulation were striking: when the filter cutoffs were set to 3900-5400 Hz, so that the components were unresolved by the peripheral auditory system, the target sounded distinctly aperiodic: a common description was that it "crackled". In contrast, when the filter cutoffs were set to 20 and 1420 Hz, so that the components of each sound were resolvable, listeners heard the second "target" sound as having a clear pitch. These initial impressions provided tentative evidence that, when two sets of unresolved harmonics are mixed within a single frequency region, listeners are in fact unable to extract the F0s of the two sounds. The main aim of the experiments described here was to test this preliminary conclusion.

Experiment 1 also investigated a potential strategy which listeners might use to extract the F0 of a target complex embedded in a second harmonic complex. It was motivated by the observations that the performance of ΔF0-based speech-separation algorithms is greatly improved by advance knowledge of the F0 of at least one of the sources (Parsons, 1976; Stubbs and Summerfield, 1988), and that similar advantages can be gained by models of human performance (Assmann and Summerfield, 1990). We wondered whether listeners could obtain this information by exploiting the onset asynchronies which often occur between competing sources: specifically, when a "masking" complex is turned on before a target complex, can listeners estimate the masker's F0 from that portion that occurs before the target, and use this information to extract the target's F0? We therefore compared the accuracy with which listeners could extract the F0 of a target complex, in conditions where it was mixed with a synchronous and with an asynchronous masker.

A. Experiment 1

1. Method

In all conditions, listeners had to identify which of two sequentially presented 200-ms harmonic complex tones (the "targets"), each of which had been bandpass-filtered, had the higher F0. The bandpass filters were set to either 20-1420 Hz (the "low frequency region") or to 3900-5400 Hz (the "high frequency region") The F0s of the two targets differed by an amount ΔF0, and were geometrically centered on 210 Hz. The targets were presented either in isolation (except for the presence of a continuous low-level pink noise), or were mixed with a masker consisting of the first 39 harmonics of 210 Hz, filtered identically with the targets, and presented at a level of 45 dB/component in the filter passband. When the masker was present, it could be either gated synchronously with the targets, or was turned on 150 ms before and off 150 ms after each target . Sensitivity (d') was measured for each combination of two frequency regions and three masker conditions, for seven ΔF0s ranging from 0.5% to 34.5% of 210 Hz.

For the low-frequency, resolved group of harmonics, it was necessary to ensure that listeners were in fact extracting the F0 of the target in each interval, rather than basing their discrimination on the frequency of one of its harmonics (Faulkner, 1985; but see Moore and Glasberg, 1990) Therefore, in this region, the target consisted of a

different subset of harmonics in the two intervals of each trial: in one interval harmonics 1,4,5,7,10,11,12... were present, whereas the other interval contained harmonics 2,3,6,8,9,12... The target was presented at a level of 45 dB/component. In the high-frequency region, no harmonics were missing from the target, but its SPL was adjusted separately for each listener on the basis of a preliminary masking experiment, so that its sensation level in the presence of the masker was the same as that of a low-frequency target. On average, this meant that the high-frequency target was presented at a level of about 48 dB/component.

2. Results

The general pattern of results was very similar across the three listeners who took part, and so mean data are presented in Fig.1. For the low frequency region (left-hand panel), the generally good performance in the "quiet" condition (triangles) was only slightly degraded by the presence of either a synchronous or an asynchronous masker. The fact that sensitivity was similar with the two masker types means that we have no evidence that presenting portions of the masker before and after the target, thereby providing the listener with prior knowledge of its F0, helps listeners to extract the F0 of the target.

In the high-frequency region, presenting an asynchronous masker (squares) produced a dramatic drop in sensitivity, which was much greater than the drop seen in the corresponding low-frequency condition. This finding is consistent with the aperiodic "crackle" or "noise-like" percept of the target in this condition, and with our preliminary conclusion that, when two groups of unresolved harmonics are mixed within the same frequency region, listeners are unable effectively to extract their F0s. It seems to be related to resolvability rather than to frequency region *per se* : in another experiment, we observed a similar large deterioration in the low region when the F0 was reduced from 210 Hz to 62.5 Hz.

In contrast, the effect of the synchronous masker (circles) was at odds with listeners' percepts: although it too produced a "crackle" percept when mixed with the target, it resulted in only a small drop in sensitivity, which was similar to that seen in the low-frequency region, where two clear pitches could be heard in the target/masker mixture.

B. Interim discussion

Why is it that the synchronous masker had only a modest effect in both frequency regions, but that the asynchronous masker produced a much larger deterioration which was specific to the high frequency region? Although we have not formulated any precise theories to account for this finding, it seems important to distinguish between two general classes of explanation. The first of these is that, even in the high-frequency region, listeners *can* extract the F0 of the target complex from that of the masker, but that some aspect of the asynchronous masker prevents listeners from performing an accurate comparison between the two intervals. For example, in the first interval of each trial, the relatively strong pitch of the trailing portion of the masker might interfere with the encoding and/or processing of the preceeding target's F0. Although no evidence for such a phenomenon has previously been reported for stimuli such as ours, there is evidence that such "retrospective" effects can impair the frequency discrimination of brief sinusoidal targets (Massaro, 1975; Kelly and Watson, 1986). However, this class of explanation does not account for the fact that the gating the masker asynchronously did not degrade performance in the low (resolved) frequency region. An alternative explanation, consistent with our preliminary interpretation, is that, in the high-frequency region, listeners *cannot* extract the F0s of both the target and masker, but that gating the two complexes synchronously allowed listeners to perform the task by some other means. One way in which this could happen would occur if gating the masker and target synchronously caused them to become perceptually fused, and allowed listeners to base discrimination on the average F0 of

the masker and target, which differed between the two intervals of each trial. The next experiment tested between these two general classes of explanation.

C. Experiment 2

1. Rationale and Method

If the trailing portion of the asynchronous masker interferes with the processing of the target, then it should have no effect in a task which does not require its F0 to be compared to that of a subsequent stimulus. Experiment 2 therefore measured sensitivity to F0 differences between two simultaneous targets, one in the low and one in the high frequency region. The task was performed both in quiet (except for a continuous pink noise), and in the presence of a continuous harmonic masker which could be either in the low or in the high region (continuous rather than asynchronous maskers were used because they made the task slightly easier to explain and to perform). The rationale was that, if listeners can extract the F0s of two groups of (resolved) components in the low-frequency region, but not those of two (unresolved) groups in the high-frequency region, then they should perform above chance only when the masker is in the low region: in this condition, the F0 of the "low" target can be extracted from the masker and compared to that of the uncorrupted "high" group. If, however, listeners can extract two intermingled F0s in both regions, then performance in this simultaneous task should not depend strongly on which region contains the masker. Again, the low- and high- frequency targets were presented at equal sensation levels.

2. Results

Fig. 2 shows psychometric functions, averaged across four listeners, for the detection of simultaneous ΔF0s. One aspect of the data is that, even in the absence of a masker and with a ΔF0 of 34.5%, performance corresponded to a d' of only about 1.5. This modest level of sensitivity is typical for the detection of ΔF0s between a resolved and an unresolved group of harmonics, particularly when, as here, the two groups are well separated in frequency (Carlyon, 1992; Carlyon and Shackleton, 1994). The second, and most important, finding is that the high-frequency masker produced a much greater drop in sensitivity than did the low-frequency masker. As in experiment 1 and in our preliminary observations, the high-frequency masker caused the high-frequency complex to sound distinctly aperiodic - listeners would spontaneously complain to that experimenter that " I don't like those conditions where you turn a burst of noise on with the tone"! The results of this experiment, then, supports our initial conclusion that listeners are very poor at extracting the F0s of two groups of unresolved harmonics which occupy the same frequency region. In contrast, they are relatively good at extracting the F0s of two groups of resolved harmonics.

III. SUMMARY AND CONCLUSIONS

(i) Listeners can compare F0s between two groups of harmonics occupying different frequency regions. They are better at doing so when both groups are resolved or are both unresolved by the peripheral auditory system, than when the two groups differ in resolvability. Carlyon and Shackleton (1994) have interepreted this as evidence that different mechanisms encode the F0s of resolved and of unresolved harmonics.

(ii) When a "target" group of harmonics is mixed into the same frequency region as a masker group with a different F0 , listeners can extract the target's F0 when the two sets of harmonic are resolvable by the peripheral auditory system. Presumably, in this case, there are auditory filters whose outputs are dominated by individual harmonics of the target, and listeners can combine the information flowing from these filters to estimate the target's F0. In contrast, when the two groups are unresolved, there are no auditory filters whose output is dominated either by individual harmonics of the target

or by its envelope. In this condition, listeners are unable to extract the F0s of the target and masker, as the relevant information is combined in the outputs of the same auditory filters.

IV. REFERENCES

Assmann, P. and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680-697.

Brokx, J. P. L. and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23-36.

Carlyon, R. P. (**1992**). "Detecting F0 differences and pitch-pulse asynchronies," in *The Auditory Processing of Speech. From Sounds to Words* edited by B. Schouten (Mouton-De Gruyter., Berlin), pp. 149-156.

Carlyon, R. P., Demany, L. and Semal, C. (**1992**). "Detection of across-frequency differences in fundamental frequency," J. Acoust. Soc. Am. **91**, 279-292.

Carlyon, R. P. and Moore, B. C. J. (**1984**). "Intensity discrimination: A severe departure from Weber's law," J. Acoust. Soc. Am. **76**, 1369-1376.

Carlyon, R. P. and Shackleton, T. M. (**1994**). "Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms?," J. Acoust. Soc. Am. **95**, 3541-3554.

Cooke, M. (**1992**). "Modelling sound source separation," in *The Auditory Processing of Speech. From Sounds to Words* edited by B. Schouten (Mouton-De Gruyter., Berlin), pp. 149-156.

Faulkner, A. (**1985**). "Pitch discrimination of harmonic complex signals: Residue pitch or multiple component discrimination," J. Acoust. Soc. Amer. **78**, 1993-2004.

Kelly, W. J. and Watson, C. S. (**1986**). "Stimulus-based limitations on the discrimination between different temporal orders of tones," J. Acoust. Soc. Am. **79**, 1934-1938.

Massaro, D. W. (**1975**). "Backward recognition masking," J. Acoust. Soc. Amer. **58**, 1059-1065.

Meddis, R. and Hewitt, M. (**1991**). "Virtual pitch and phase sensitivity studied using a computer model of the auditory periphery: Pitch identification," J. Acoust. Soc. Am. **89**, 2866-2882.

Meddis, R. and Hewitt, M. (**1992**). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. **91**, 233-245. submitted.

Moore, B. C. J. and Glasberg, B. R. (**1990**). "Frequency discrimination of complex tones with overlapping and non-overlapping harmonics," J. Acoust. Soc. Am. **87**, 2163-2177.

Parsons, T. W. (**1976**). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. **60**, 911-918.

Patterson, R., D, Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand, M. (**1991**). "Complex sounds and auditory images," in *Auditory*

Physiology and Perception edited by Y. Cazals, L. Demany and K. Horner (Pergamon, Oxford), pp. 429-446.

Scheffers, M. T. M. (**1983**). "Sifting vowels: auditory pitch analysis and sound segregation," Doctoral dissertation. Univ. Groningen, Netherlands.

Slaney, M. and Lyon, R. F. (**1990**). "A perceptual pitch detector," Proc. Int. Conf. Acoustics, Speech, and Signal Processing 357-360.

Stubbs, R. J. a. and Summerfield, A. Q. (**1988**). "Evaluation of two voice-separation algorithms using normally-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **84**, 1236-1249.

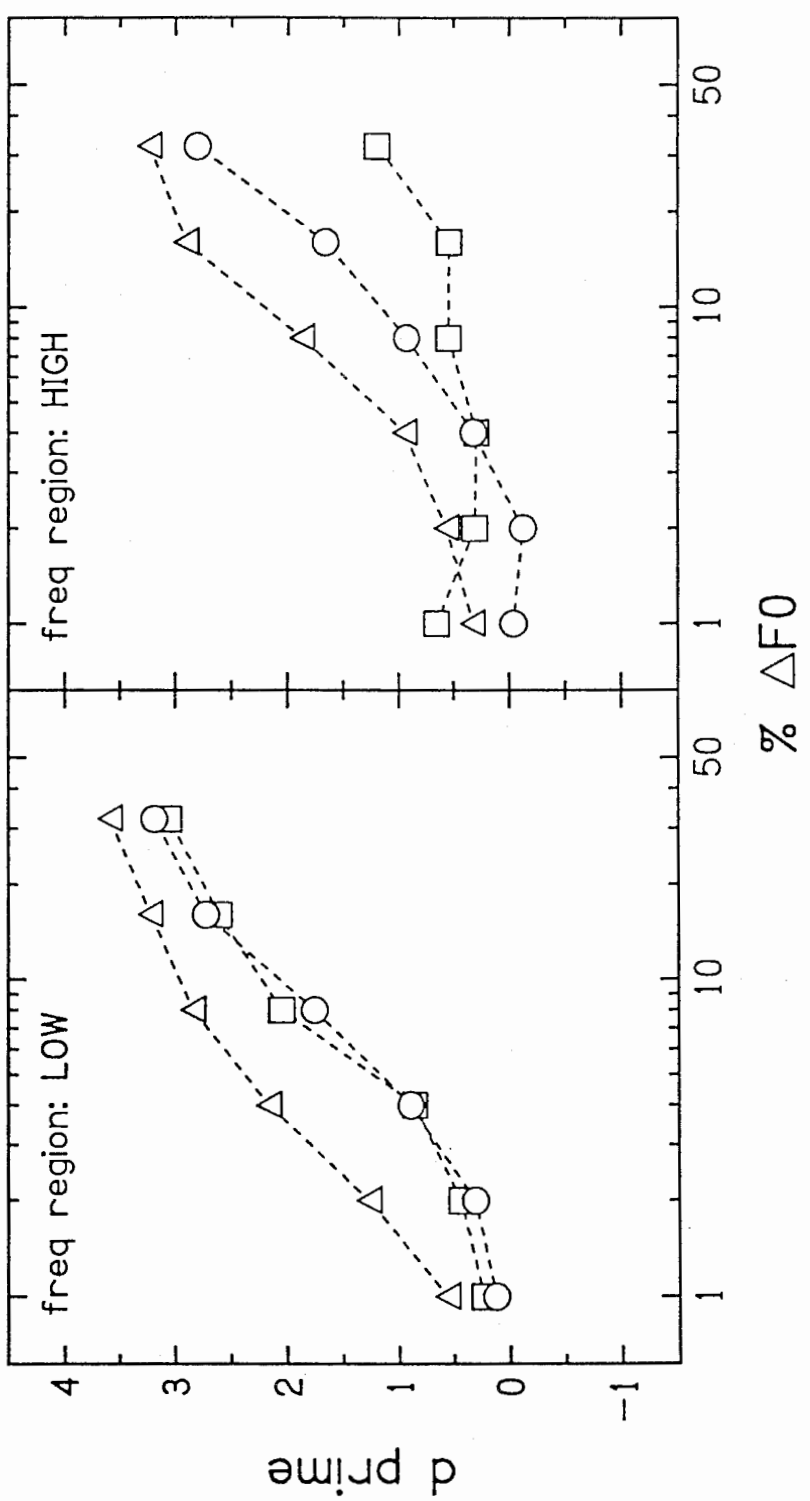

MASKER
NONE
ASYNCH
SYNCH
freq region: HIGH
freq region: LOW
d prime
% ΔF0
4
3
2
1
0
-1
1
10
50

Figure 1.

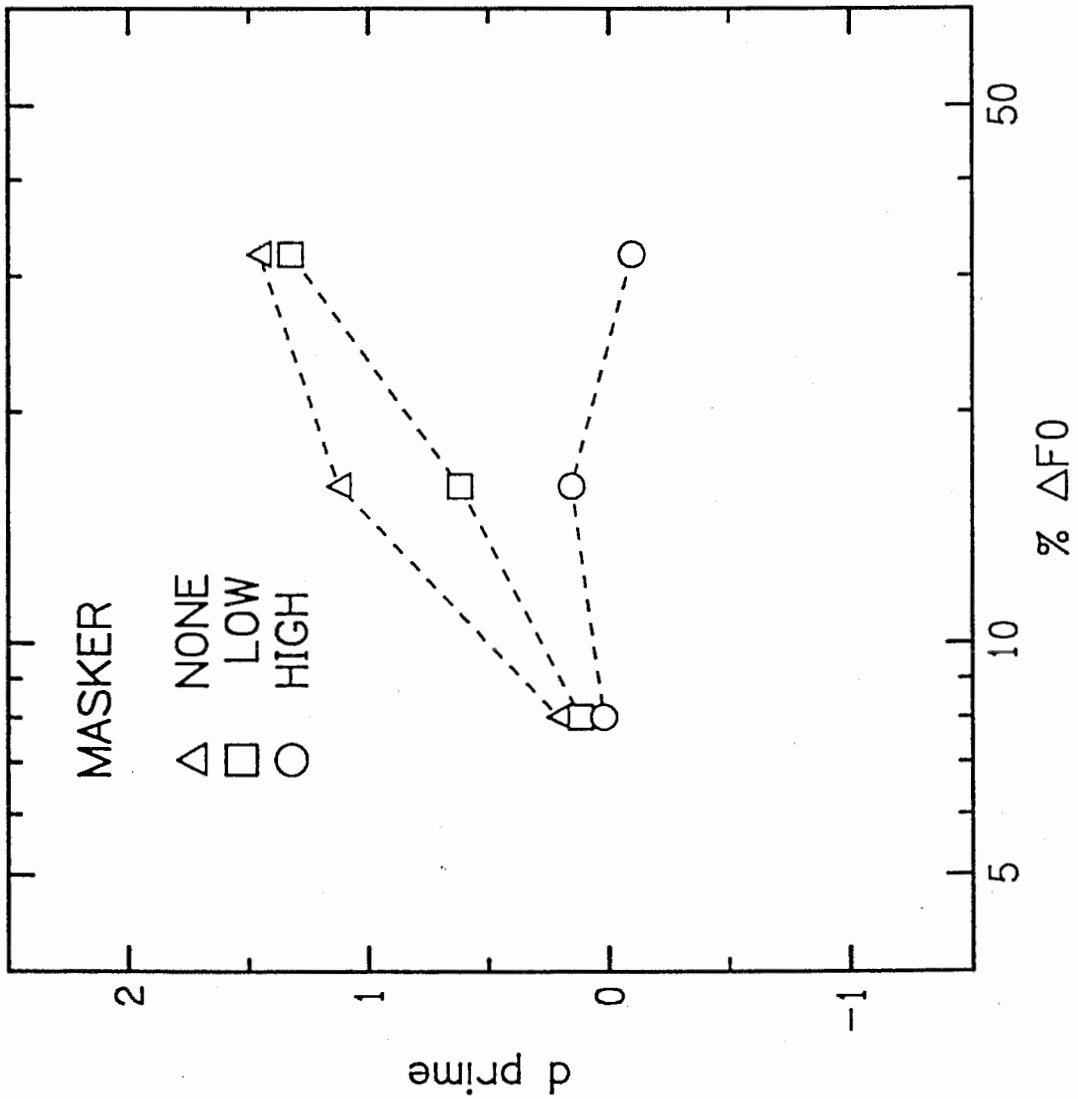


Figure 2.

AN INTRODUCTION TO AUDITORY MODEL INVERSION

Malcolm Slaney
Interval Research Corporation
1801-C Page Mill Road, Palo Alto, CA 94304
malcolm@interval.com

1 – INTRODUCTION[1]

My interest in auditory models and perceptual displays [2] is motivated by the problem of sound understanding, especially the separation of speech from noisy backgrounds and interfering speakers. The correlogram and related representations are a pattern space within which sounds can be "understood" and "separated" [3][4]. I am therefore interested in resynthesizing sounds from these representations as a way to test and evaluate sound separation algorithms, and as a way to apply sound separation to problems such as speech enhancement. The conversion of sound to a correlogram involves the intermediate representation of a cochleagram, as shown in Figure 1, so cochlear-model inversion is addressed as a one piece of the overall problem.

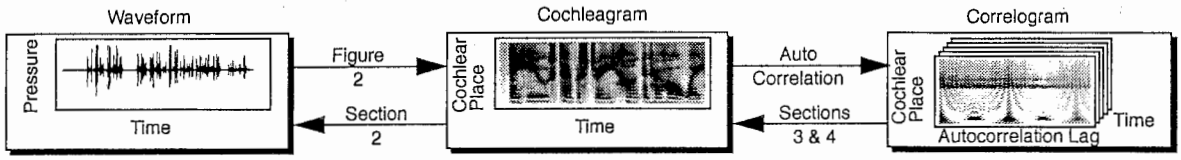


Figure 1. Three stages in low-level auditory perception are shown here. Sound waves are converted into a detailed representation with broad spectral bands, known as cochleagrams. The correlogram then summarizes the periodicities in the cochleagram with short-time autocorrelation. The result is a perceptual movie synchronized to the acoustic signal. The two inversion problems addressed in this work are indicated with arrows from right to left.

The inversion techniques described here are important because they allow us to readily evaluate the results of sound separation models that "zero out" unwanted portions of the signal in the correlogram domain. This work extends the convex projection approach of Irino [5] and Yang [6] by considering a different cochlear model, and by including the correlogram inversion. The convex projection approach is well suited to "filling in" missing information. While this paper only describes the process for one particular auditory model, the techniques are equally useful for other models.

This paper describes three aspects of the problem: cochleagram inversion, conversion of the correlogram into spectrograms, and spectrogram inversion. A number of reconstruction options are explored in this paper. Some are fast, while other techniques use time-consuming iterations to produce reconstructions perceptually equivalent to the original sound. Fast versions of these algorithms could allow us to separate a speaker's voice from the background noise in real time.

2 – COCHLEAGRAM INVERSION

Figure 2 shows a block diagram of the cochlear model [7] that is used in this work. The basis of the model is a bank of filters, implemented as a cascade of low-pass filters, that splits the input signal into broad spectral bands. The output from each filter in the bank is called a channel. The energy in each channel is detected and used to adjust the channel gain, implementing a simple model of auditory sensitivity adaptation, or automatic gain control (AGC). The half-wave rectifier (HWR) detection nonlinearity provides a waveform for each channel that roughly represents the instantaneous neural firing rate at each position along the cochlea.

The cochleagram is converted back into sound by reversing the three steps shown in Figure 2. First the AGC is divided out, then the negative portions of each cochlear channel are recovered by using the fact that each channel is spectrally limited. Finally, the cochlear filters are inverted by running the filters backwards, and then correcting the resulting spectral slope.

---

1. This work was performed by Malcolm Slaney, Daniel Naar and Richard F. Lyon while all three were employed at Apple Computer. The mathematical details of this work were presented at the 1994 ICASSP[1].

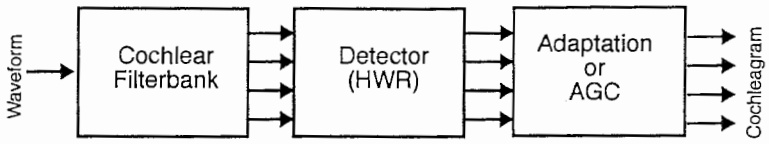


Figure 2. Three stages of the simple cochlear model used in this paper are shown above.

The AGC stage in this cochlear model is controlled by its own output. It is a combination of a multiplicative gain and a simple first-order filter to track the history of the output signal. Since the controlling signal is directly available, the AGC can be inverted by tracking the output history and then dividing instead of multiplying. The performance of this algorithm is described by Naar [8] and will not be addressed here. It is worth noting that AGC inversion becomes more difficult as the level of the input signal is raised, resulting in more compression in the forward path.

The next stage in the inversion process can be done in one of two ways. After AGC inversion, both the positive values of the signal and the spectral extant of the signal are known. Projections onto convex sets [9], in this case defined by the positive values of the detector output and the spectral extant of the cochlear filters, can be used to find the original signal. This is shown in the left half of Figure 3. Alternatively, the spectral projection filter can be combined with the next stage of processing to make the algorithm more efficient. The increased efficiency is due to better match between the spectral projection and the cochlear filterbank, and due to the simplified computations within each iteration. This is shown in the right half of Figure 3. The result is an algorithm that produces nearly perfect results with no iterations at all.

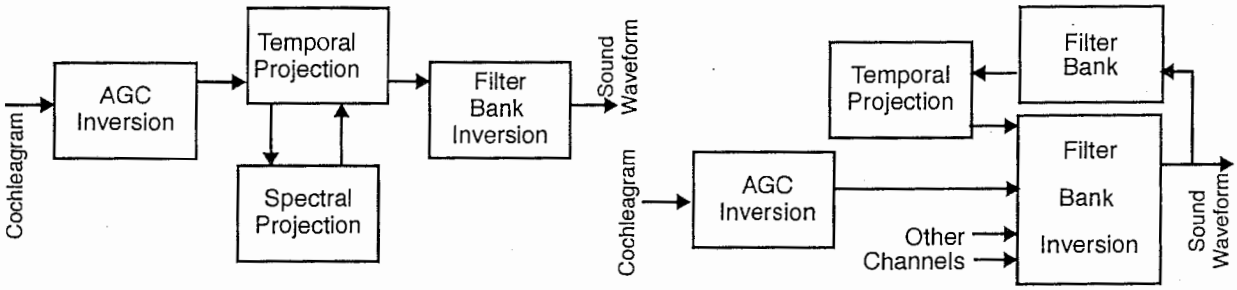


Figure 3. There are two ways to use convex projections to recover the information lost by the detectors. The conventional approach is shown on the left. The right figure shows a more efficient approach where the spectral projection has been combined with the filterbank inversion

Finally, the multiple outputs from the cochlear filterbank are converted back into a single waveform by correcting the phase and summing all channels. In the ideal case, each cochlear channel contains a unique portion of the spectral energy, but with a bit of phase delay and amplitude change. For example, if we run the signal through the same filter the spectral content does not change much but both the phase delay and amplitude change will be doubled. More interestingly, if we run the signal through the filter backwards, the forward and backward phase changes cancel out. After this phase correction, we can sum all channels and get back the original waveform, with a bit of spectral coloration. The spectral coloration or tilt can be fixed with a simple filter. A more efficient approach to correct the spectral tilt is to scale each channel by an appropriate weight before summing, as shown in Figure 4. The result is a perfect reconstruction, over those frequencies where the cochlear filters are non-zero.

Figure 4. Two approaches are shown here to invert the filterbank. The left diagram shows the normal approach, the right figure shows a more efficient approach where the spectral-tilt filter is converted to a simple multiplication.

Figure 5 shows results from the cochleagram inversion procedure. An impulse is shown on the left, before and after 10 iterations of the HWR inversion (using the algorithm on the right half of Figure 3). With no iterations the result is nearly perfect, except for a bit of noise near the center. The overall curvature of the baseline is due to the fact that information near DC has been lost as it travels through the auditory system and there is no way to recover it with the information that we have. A more interesting example is shown on the right. Here the word "tap"[1] has been reconstructed, with and without the AGC inversion. With the AGC inversion the result is nearly identical to the original. The auditory system is very sensitive to onsets and quickly adapts to steady state sounds like vowels. It is interesting to compare this to the reconstruction without AGC inversion. Without the AGC, the result is similar to what the ear hears, the onsets are more prominent and the vowels are deemphasized. This is shown in the right half of Figure 5.

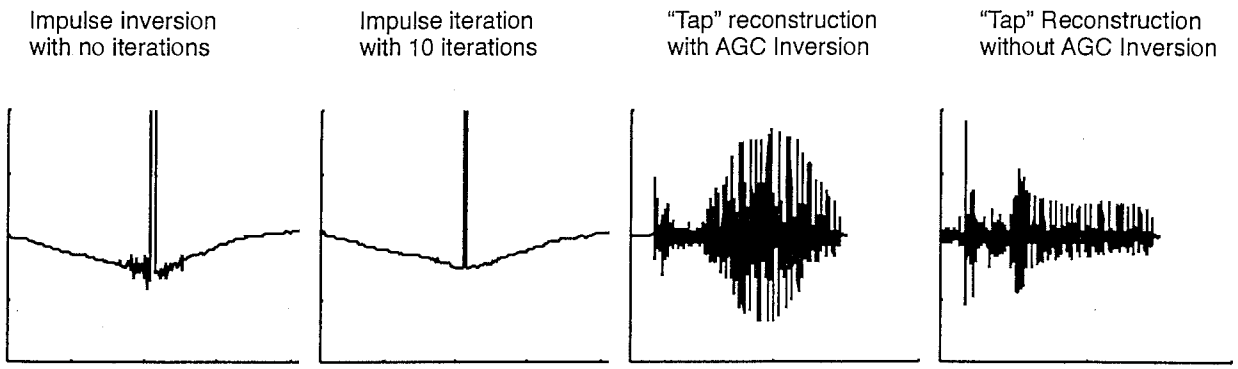


Figure 5. The cochlear reconstructions of an impulse and the word "tap" are shown here. The first and second reconstructions show an impulse reconstruction with and without iterations. The third and fourth waveforms are the word "tap" with and without the AGC inversion.

3 – CORRELOGRAM INVERSION

The correlogram is an efficient way to capture the short-time periodicities in the auditory signal. Many mechanical measurements of the cochlea have shown that the response is highly non-linear. As the signal level changes there are large variations in the bandwidth and center frequency of the cochlear response. With these kinds of changes, it is difficult to imagine a system that can make sense of the spectral profile. This is especially true for decisions like pitch determination and sound separation.

But through all these changes in the cochlear filters, the timing information in the signal is preserved. The spectral profile, as measured by the cochlea, might change, but the rate of glottal pulses is preserved. Thus I believe the auditory system is based on a representation of sound that makes short-time periodicities apparent. One such representation is the correlogram. The correlogram measures the temporal correlation within each channel, either using FFTs which are most efficient in computer implementations, or neural delay lines much like those found in the binaural system of the owl.

The process of inverting the correlogram is simplified by noting that each autocorrelation is related by the Fourier transform to a power spectrum. By combining many power spectrums into a picture, the result is a spectrogram. This process is shown in Figure 6. In this way, a separate spectrogram is created for each channel. There are known techniques for converting a spectrogram, which has amplitude information but no phase information, back into the original waveform. The process of converting from a spectrogram back into a waveform is described in Section 4. The correlogram inversion process consists of inverting many spectrograms to form an estimate of a cochleagram. The cochleagram is inverted using the techniques described in Section 2.

One important improvement to the basic method is possible due to the special characteristics of the correlogram. The essence of the spectrogram inversion problem is to recover the phase information that has been thrown away. This is an iterative procedure and would be costly if it had to be performed on each channel. Fortunately, there is quite a bit of overlap between cochlear channels. Thus the phase recovered from one channel can be used to initialize the spec-

1.The syllable "tap", samples 14000 through 17000 of the "train/dr5/fcdf1/sx106/sx106.adc" utterance on the TIMIT Speech Database, is used in all voiced examples in this paper.

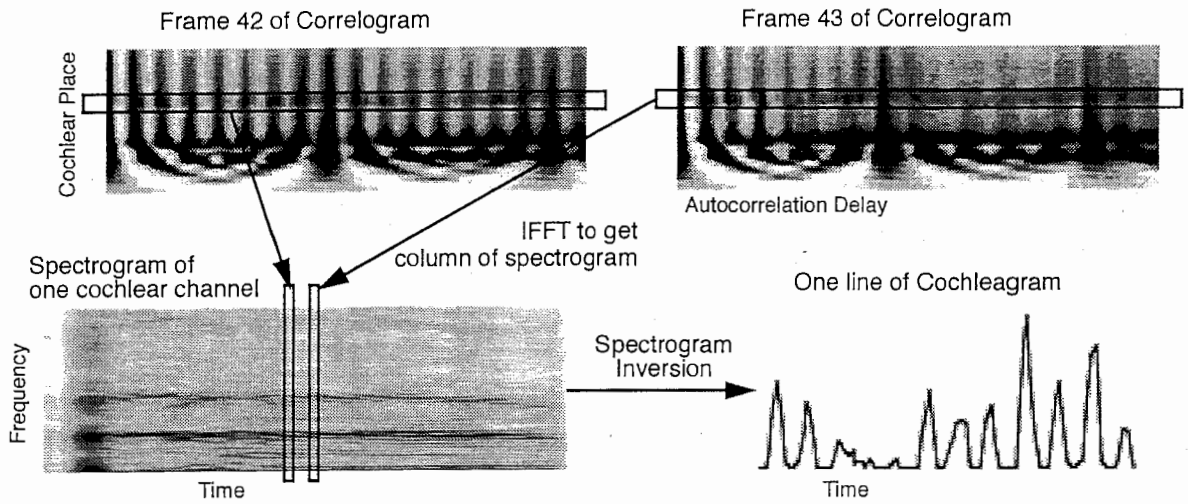


Figure 6. Correlogram inversion is possible by noting that each row of the correlogram contains the same information as a spectrogram of the same row of cochleagram output. By converting the correlogram into many spectrograms, the spectrogram inversion techniques described in Section 4 can be used. The lower horizontal stripe in the spectrogram is due to the narrow passband of the cochlear channel. Half-wave rectification of the cochlear filter output causes the upper horizontal stripes.

trogram inversion for the next channel. A difficulty with spectrogram inversion is that the absolute phase is lost. By using the phase from one channel to initialize the next, a more consistent set of cochlear channel outputs is recovered.

4 – SPECTROGRAM INVERSION

While spectrograms are not an accurate model of human perception, an implementation of a correlogram includes the calculation of many spectrograms. Mathematically, an autocorrelation calculation is similar to a spectrogram or short-time power spectrum. One column of a conventional spectrogram is related to an autocorrelation of a portion of the original waveform by a Fourier transform (see Figure 6). Unfortunately, the final representation of both spectrograms and autocorrelations is missing the phase information. The main task of a spectrogram inversion algorithm is to recover a consistent estimate of the missing phase. This process is not magical, it can only recover a signal that has the same magnitude spectrum as the original spectrogram. But the consistency constraint on the time evolution of the signal power spectrum also constrains the time evolution of the spectral phase.

The basic procedure in spectrogram inversion [10] consists of iterating between the time and the frequency domains. Starting from the frequency domain, the magnitude but not the phase is known. As an initial guess, any phase value can be used. The individual power spectra are inverse Fourier transformed and then summed to arrive at a single waveform. If the original spectrogram used overlapping windows of data, the information from adjacent windows either constructively or destructively interferes to estimate a waveform. A spectrogram of this new data is calculated, and the phase is now retained. We know the original magnitude was correct. Thus we can estimate a better spectrogram by combining the original magnitude information with the new phase information. It can be shown that each iteration will reduce the error.

Figure 7 shows an outline of steps that can be used to improve the consistency of phase estimates during the first iteration. As each portion of the waveform is added to the estimated signal, it is possible to add a linear phase so that each waveform lines up with the proceedings segments. The algorithm described in the paragraph above assumes an initial phase of zero. A more likely phase guess is to choose a phase that is consistent with the existing data. The result with no iterations is a waveform that is often closer to the original than that calculated assuming zero initial phase and ten iterations.

The total computational cost is minimized by combining these improvements with the initial phase estimates from adjacent channels of the correlogram. Thus when inverting the first channel of the correlogram, a cross-correlation is used to pick the initial phase and a few more iterations insure a consistent result. After the first channel, the phase of the proceeding channel is used to initialize the spectrogram inversion and only a few iterations are necessary to fine tune the waveform.

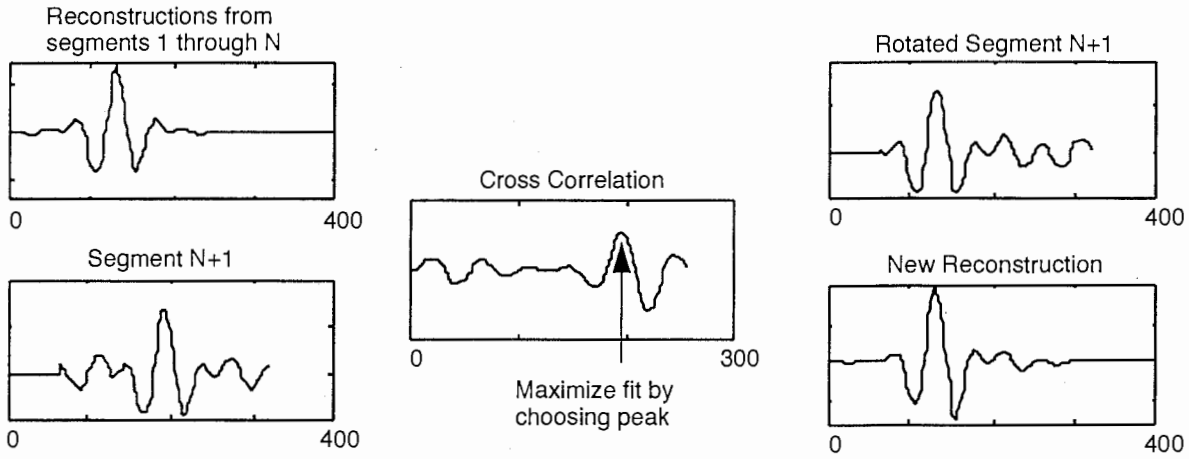


Figure 7. A procedure for adjusting the phase of new segments when inverting a spectrogram is shown above. As each new segment (bottom left) is converted from a power spectrum into a waveform, a linear phase is added to maximize the fit with the existing segments (top left.) The amount of rotation is determined by a cross correlation (middle). Adding the new segment with the proper rotation (top right) produces the new waveform (bottom right.)

5 – PUTTING IT TOGETHER

This paper has described two steps to convert a correlogram into a sound. These steps are detailed below:

1) For each row of the correlogram:
   a) Convert the autocorrelation data into power spectrum (Section 3).
   b) Use spectrogram inversion (Section 4) to convert the spectrograms into an estimate of cochlear channel output.
   c) Assemble the results of spectrogram inversion into an estimate of the cochleagram.
2) Invert the cochleagram using the techniques described in Section 2.

This process is diagrammed in Figures 1 and 6.

6 – RESULTS

Figure 8 shows the results of the complete reconstruction process for a 200Hz impulse train and the word "tap." In both cases, no iterations were performed for either the spectrogram or filterbank inversion. More iterations reduce the spectral error, but do not make the graphs look better or change the perceptual quality much. It is worth noting that the "tap" reconstruction from a correlogram looks similar to the cochleagram reconstruction without the AGC (see Figure 5.) Reducing the level of the input signal, thus reducing the amount of compression performed by the AGC, results in a correlogram reconstruction similar to the original waveform.

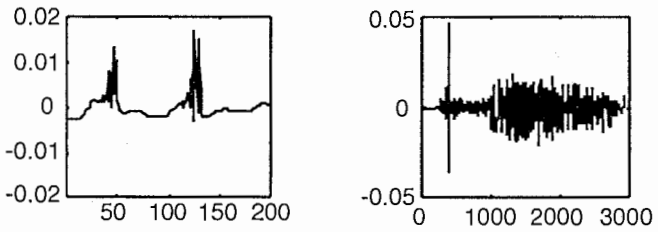


Figure 8. Reconstructions from the correlogram representation of an impulse train and the word "tap" are shown above. Reducing the input signal level, thus minimizing the effect of errors when inverting the AGC, produces results identical to the original "tap."

It is important to note that the algorithms described in this paper are designed to minimize the error in the mean-square sense. This is a convenient mathematical definition, but it doesn't always correlate with human perception. A trivial example of this is possible by comparing a waveform and a copy of the waveform delayed by 10ms. Using the mean-squared error, the numerical error is very high yet the two waveforms are perceptually equivalent. Despite this, the results of these algorithms based on mean-square error do sound good.

7 – CONCLUSIONS

This paper has described several techniques that allow several stages of an auditory model to be converted back into sound. By converting each row of the correlogram into a spectrogram, the spectrogram inversion techniques of Section 4 can be used. The special characteristics of a correlogram described in Section 3 are used to make the calculation more efficient. Finally, the cochlear filterbank can be inverted to recover the original waveform. The results are waveforms, perceptually identical to the original waveforms.

These techniques will be especially useful as part of a sound separation system. I do not believe that our auditory system resynthesizes partial waveforms from the auditory scene. Yet, all systems generate separated sounds so that we can more easily perceive their success. More work is still needed to fine-tune these algorithm and to investigate the ability to reconstruct sounds from partial correlograms.

8 – ACKNOWLEDGEMENTS

I am grateful for the inspiration provided by Frank Cooper's work in the early 1950's on pattern playback[11][12]. His work demonstrated that it was possible to convert a spectrogram, painted onto clear plastic, into sound.

This work in this paper was performed with Daniel Naar and Richard F. Lyon. We are grateful for the help we have received from Richard Duda (San Jose State), Shihab Shamma (U. of Maryland), Jim Boyles (The MathWorks) and Michele Covell (Interval Research).

9 – REFERENCES

[1] Malcolm Slaney, D. Naar, R. F. Lyon, "Auditory model inversion for sound separation," *Proc. of IEEE ICASSP*, Volume II, pp. 77-80, 1994.
[2] M. Slaney and R. F. Lyon, "On the importance of time—A temporal representation of sound," in *Visual Representations of Speech Signals*, eds. M. Cooke, S. Beet, and M. Crawford, J. Wiley and Sons, Sussex, England, 1993.
[3] R. F. Lyon, "A computational model of binaural localization and separation," *Proc. of IEEE ICASSP*, 1148-1151, 1983.
[4] M. Weintraub, "The GRASP sound separation system," *Proc. of IEEE ICASSP*, pp. 18A.6.1-18A.6.4, 1984.
[5] T. Irino, H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," *IEEE Trans. on Signal Processing*, 41, 3549-3554, Dec. 1993.
[6] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. on Information Theory*, 38, 824-839, 1992.
[7] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," *Proc. of the IEEE ICASSP*, 1282-1285, 1982.
[8] D. Naar, "Sound resynthesis from a correlogram," San Jose State University, Department of Electrical Engineering, Technical Report #3, May 1993.
[9] R. W. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," *IEEE Trans. Circuits Sys.*, vol. 22, 735, 1975.
[10] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 32, 236-242, 1984.
[11] F. S. Cooper, "Some Instrumental Aids to Research on Speech," *Report on the Fourth Annual Round Table Meeting on Linguistics and Language Teaching*, Georgetown University Press, 46-53, 1953.
[12] F. S. Cooper, "Acoustics in human communications: Evolving ideas about the nature of speech," J. Acoust. Soc. Am., 68(1), 18-21, July 1980.

The computation of loudness in the auditory continuity phenomenon

S. McAdams[1,2], M.-C. Botte[1], F. Banide[1], X. Durot[1] & C. Drake[1]

[1] Laboratoire de Psychologie Expérimentale (CNRS), Université René Descartes, EPHE, 28 rue Serpente, F-75006 Paris, France.
[2] IRCAM, 1 place Igor-Stravinsky, F-75004 Paris, France. email: smc@ircam.fr

INTRODUCTION

To recover a veridical representation of the acoustic environment, the auditory system needs to group together acoustic components that are likely to have originated from the same source into coherent mental descriptions (variously referred to as auditory "streams", "objects", "images" or "entities"). Once the streams are organized, the auditory system can compute the perceptual attributes (loudness, pitch, timbre, etc.) of the events belonging to each stream. Our experiments aimed to measure the effect of auditory organization on the computation of loudness.

Consider the stimulus sequence in Fig. 1 in which a pure tone or a noise signal alternates between two levels A and B. One might imagine at first glance that a listener would hear an alternating sequence of loud and soft tones or noise bursts, or a single sound stream changing periodically in level. If we asked listeners to adjust a comparison sequence to the loud or soft parts of the test sequence, we should obtain matches in the vicinity of A and B, respectively, with perhaps some modifications due to temporal masking effects and the time course of energy summation within each burst and loudness summation across bursts. Such a prediction would be made by classic time-varying loudness models which do not consider the effect of temporally overlapping sound signals originating from separate sources.

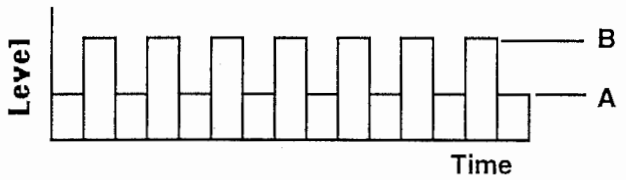


Figure 1. Stimulus sequence alternating between levels A and B.

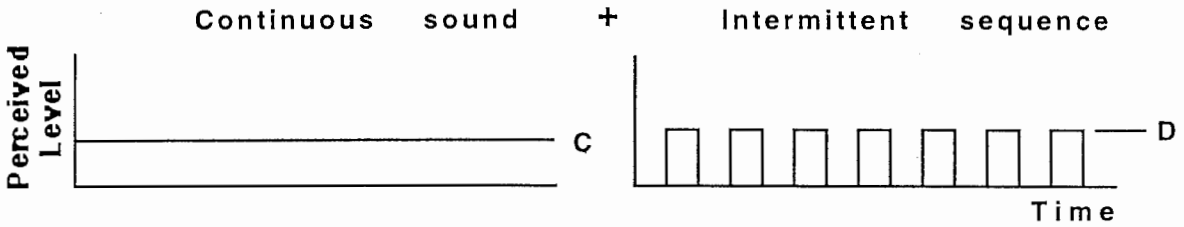


Figure 2. Percepts resulting from the alternating sequence in Fig. 1: a continuous sound of level C and a sequence of intermittent bursts of level D.

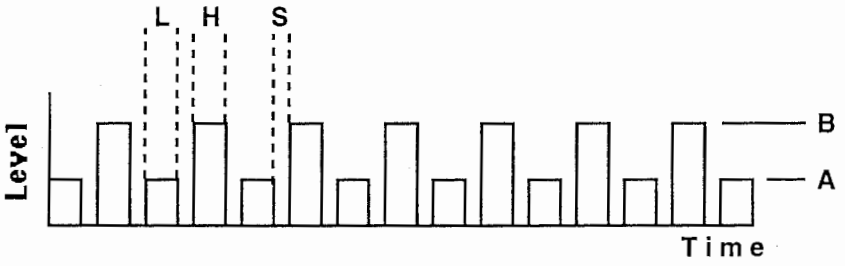
However, van Noorden (1975) and others have shown that one hears continuity under certain conditions that depend on the difference between levels A and B, on the presence of silences between bursts, etc. (Thurlow, 1957; Houtgast, 1972). Warren, Obusek & Ackroff (1972), for example, demonstrated that such a stimulus, with a level difference of between about 3 and 10 dB, generally gives rise to a perception not of alternation but of an intermittent tone "superimposed" on a continuous tone (Fig. 2). In Bregman's (1990) conception of this phenomenon, the high-level part is perceptually split into two simultaneous components, one being assigned to the continuous tone and the other to the intermittent tone. This phenomenon has been called "auditory continuity" (Thurlow & Elfner, 1959) since the low-level signal is heard to continue "through" the intermittent signal or "auditory induction" (Warren *et al.*, 1972) since the high-level part induces a perception of continuity of the low-level part. Similar types of phenomena have been demonstrated for speech interrupted by noise, in which a continuous speech signal appears to be "perceptually restored" during the noise burst. In these cases, the auditory system would appear to have interpreted the signal as being composed of a continuous sound upon which another signal is superimposed at regular intervals. What is of interest in this phenomenon is what it might tell us about how the auditory system disentangles the respective perceptual attributes of superimposed signals.

Paper presented at the ATR Workshop on "A Biological Framework for Speech Perception and Production", Kyoto, 16-17 September 1994

Bregman's (1990) "old-plus-new" strategy makes some predictions about this kind of perceptual phenomenon. This strategy is hypothesized to perform an interpolation (perceptual restoration) between the properties of the softer sounds occurring before and after the louder interrupting sound, but only if the auditory information indicates that the soft sound could have been present during the occurrence of the loud sound. Subsequently, the signal in the time interval occupied by the loud sound is interpreted as resulting from a mixture of the soft (old) and loud (new) sounds. The computation of the loudness of the intermittent stream would be based on a subtraction of the level of the restored part of the continuous sound from the global level of the intense part of the sequence. If we asked listeners to adjust a comparison sequence to the continuous and intermittent parts of this test sequence, we should obtain a level C (Fig. 2) in the vicinity of A (Fig. 1) for the continuous part and a level D (Fig. 2) for the intermittent part that would depend on the subtraction law the auditory system used to derive the loudness of this latter part. If identical stimuli added in phase are presented to one or both ears, one might expect a law computed on acoustic pressure ($D=B-A$), whereas with stimuli the phase relations of which are indeterminate (as with signals of unknown properties, or even known signals presented in a reverberant environment), one might expect a law computed on acoustic power ($D^2=B^2-A^2$). In both cases, however, adjusted level D would be less than the physical level B.

GENERAL METHOD

To test this hypothesis, we presented to listeners sequences that alternated between a high level (H) and a low level (L) as in Fig. 3. Listeners were asked to adjust the level of a comparison stimulus so that its loudness matched that of a specific part of the test stimulus that varied with the experiment or within a block of trials. Stimulus parameters were varied to test the dependence of adjusted levels C and D (Fig. 2) on physical levels A and B (Fig. 1) under conditions in which listeners either clearly experienced the auditory continuity illusion or could not hear it. Control experiments were also performed to test the precision of level adjustments in the absence of auditory continuity.



H/L	Exp. 1-3	: {200/200, 100/100, 100/300, 100/700}
	Exp. 4	: 100/300 ms
S	Exp. 1-3	: 0 ms
	Exp. 4	: {30, 100} ms
A	60 ± {1,3,5} dB SPL	
B	A + {2, 6, 10} dB	

Figure 3. General stimulus conditions used for test sequences.

Test sequences were composed of one of two types of stimulus bursts: 1 kHz pure tone (PT) or 140 Hz noise band centered on 1 kHz (NB). They were also presented with four different duty cycles to study effects of loudness summation over time (H duration/L duration in ms: 200/200, 100/100, 100/300, 100/700). The level of L bursts was varied randomly within 5 dB of 60 dB SPL and the level of H bursts was either 2, 6 or 10 dB greater than that of L. The prediction was that the perceived loudness of the intermittent part should vary systematically with this level difference, always being adjusted to a level below that of the high-level part of the stimulus sequence. For a mechanism operating on acoustic pressure, listeners perceiving the illusion should adjust an intermittent comparison sequence to levels that are 13.7, 6.0 and 3.3 dB, respectively, below the level of B. For a mechanism operating on acoustic power, the adjusted levels should be 4.3, 1.3, and 0.5 dB, respectively, below B.

Individual bursts had 5 ms amplitude ramps. The H and L burst onsets and offsets either overlapped by 2.5 ms (0 ms silence) or were separated by 30 or 100 ms silent intervals. In the case of contiguous bursts, the continuity illusion is quite strong if the sequence is presented monaurally or

diotically, but it is absent if H and L bursts are presented to separate ears. With silent intervals, the illusion is generally absent or quite weak for a 30 ms silence at the levels we used, and is almost never perceived with 100 ms silences.

For sequences producing the illusion, subjects were asked on a given trial either to adjust the level of a continuous comparison tone to match the level of what appeared to be continuous in the test sequence, or to adjust the level of an intermittent sequence to match the level of what appeared to be intermittent in the test sequence. For sequences not producing the illusion, intermittent comparison sequences of similar temporal structure were presented and the subject was asked to match either the higher or the lower level in the test sequence using ear of presentation or duration cues to focus on the target stream. The starting levels of the comparison sequence were chosen at random from ±{7, 8, 9, 10} dB relative to A. Stimuli were presented in blocks comprising a given stimulus type (PT or NB) and duty cycle. Each block was repeated five times by each of eight subjects in each experiment.

EXPERIMENT 1: Alternating sequences producing the continuity illusion—no silences.

The goal of this experiment was to test the main hypothesis that the loudness of an alternating sequence organized into a continuous and an intermittent stream is partitioned into two quantities that may be computed either on the basis of pressure or power. On each trial the alternating sequence was followed by a comparison sequence that was either continuous or intermittent, the latter having the same temporal structure as the H bursts. In a familiarization phase, all subjects reported the continuity illusion though the effect was weaker for the 2 dB difference in level between A and B, sometimes heard more as a fluctuating level. Subjects also reported that the bursts in the intermittent stream in the alternating sequence were degraded in terms of the attack quality and tone color compared with the isolated intermittent sequence (similarly to results reported by Warren, Bashford, Healy & Brubaker, 1994).

Results

The dependent variable was the adjustment "error" with respect to the physical level (C–A, D–B). There was nearly perfect adjustment of the level of the continuous stream (Fig. 4). The intermittent stream was always adjusted to a level less than B and the more so as the A–B level difference was small. In fact, D was adjusted even lower than A for a difference of 2 dB. There was no significant difference between adjusted levels for PT and NB stimuli, so the same subtraction law seems to apply to both and gives values between the predictions based on power and pressure both for our data and for those of Warren *et al.* (1994) (Fig. 4). Adjustments for the intermittent stream were progressively less precise as the A–B difference decreased, suggesting subjects' difficulty in segregating the intermittent stream for the 2 dB difference. The duty cycle had no effect on the levels adjusted for the intermittent stream.

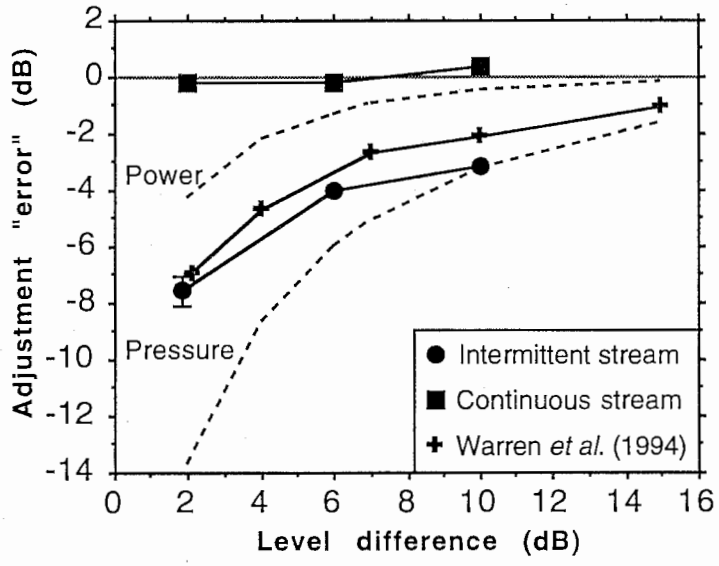


Figure 4. Experiment 1: Mean adjustment "error" as a function of the level difference between H and L bursts. Vertical bars (where visible) show standard error. Also shown are the predictions for intermittent stream adjustments according to pressure and power subtraction laws. Data from Warren *et al.* (1994) for broad-band noise (2 and 4 dB difference) and pure tones (4, 7, 10, and 15 dB) are included for comparison.

EXPERIMENT 2 : Comparison of intermittent and continuous sounds.

The goal of this control experiment was to verify whether the precision of adjustment depended on the intermittence of the test or comparison sequences. Two types of stimuli were used that resemble those in Fig. 2: a continuous (CONT) sound of 2600 ms duration and a series of intermittent (INT) sequences with 7 sounds (duration of H bursts) separated by silences (duration of L bursts). Each stimulus type could be presented either as test stimulus or as adjustable comparison stimulus: INT/INT, CONT/CONT, INT/CONT, CONT/INT. The stimuli were presented diotically.

Results

No differences in adjusted level were found between PT and NB stimuli. The adjustment precision was excellent (error of 0.1 dB on average) when the test and comparison sequences were of the same type (INT/INT, CONT/CONT). However, the level of the intermittent sequence was systematically underestimated in comparisons across sequence types (error of –2.5 dB for continuous adjusted to intermittent and of +1.8 dB for the reciprocal condition). The error increased with the difference in duration in the two parts of the duty cycle. This effect may be explained by the temporal summation of loudness.

EXPERIMENT 3 : Dichotic alternating sequences—no continuity illusion.

The aim of this control experiment was to verify the precision with which level adjustments are made with reference to a test stimulus that had the same temporal configuration as that of the sequence producing the illusion (S=0), but that did not produce the continuity illusion. Accordingly, the H and L portions of the sequence were sent to the right and left earphones, respectively. The adjustable comparison sequence consisted of a series of bursts of identical duration as those in the target ear and was presented to that ear. At the beginning of the experiment all three level differences were presented to subjects who were asked what they heard. No subject reported a sensation of continuity.

Results

The dependent variable was the "error" in adjustment of the comparison sequence relative to the physical level presented. No difference was found between noise and tone stimuli. The mean values across duty cycles varied from –0.5 to +1.3 dB. The error was nearly zero for the H sequences (right ear) and about 1 dB for the L sequences (left ear). This tendency to overestimate the level of the L sequence was greater for larger level differences. The overestimation also became progressively stronger as the duration of the L bursts decreased, being 1.9 dB for 100 ms, 1.2 dB for 200 ms and 0.5 dB for 300 and 700 ms reflecting a loudness accumulation effect over about the first 200 ms of a sound.

EXPERIMENT 4: Alternating sequences with silences between bursts—continuity illusion very weak or absent.

The aim of this control experiment was to create an experimental situation similar in structure to that of the main experiment (100/300 duty cycle only) but in which the continuity illusion was not heard. Therefore 30 or 100 ms silences were introduced to separate the low- and high-level tone bursts (Fig. 3). Two groups of eight subjects completed the two conditions. Subjects' verbal reports indicated that 30 ms silences could at times give a weak impression of the illusion, but they could also learn not to hear the illusion. For these stimuli, some subjects found it difficult to focus on one level at the beginning. The 100 ms silences never gave the illusion and presented less problems of attentional focus. To focus subjects' attention on the perception of intermittent streams in the test sequence, the adjustable comparison sequences were always intermittent and corresponded identically in temporal structure (300 ms for lower-level and 100 ms for higher-level streams) to the targeted high- or low-level part of the test sequence.

Results

There was a significant effect of level difference for adjustments to both H and L streams (Fig. 5). For H streams, adjustment errors were increasingly negative with increased difference in level between the streams, while the reverse was the case for adjustments to the L streams. It would seem, therefore, that the context effect of a sequence with alternating level increases softer sounds and decreases louder sounds. This effect is much larger than the small bias found in non-alternating sequences (Exp. 2). There was no effect of the duration of the silent interval on mean adjustment errors. The variability in adjustments was slightly lower for the 100 ms silences, perhaps due to the better destruction of the illusion than was obtained with 30 ms silences.

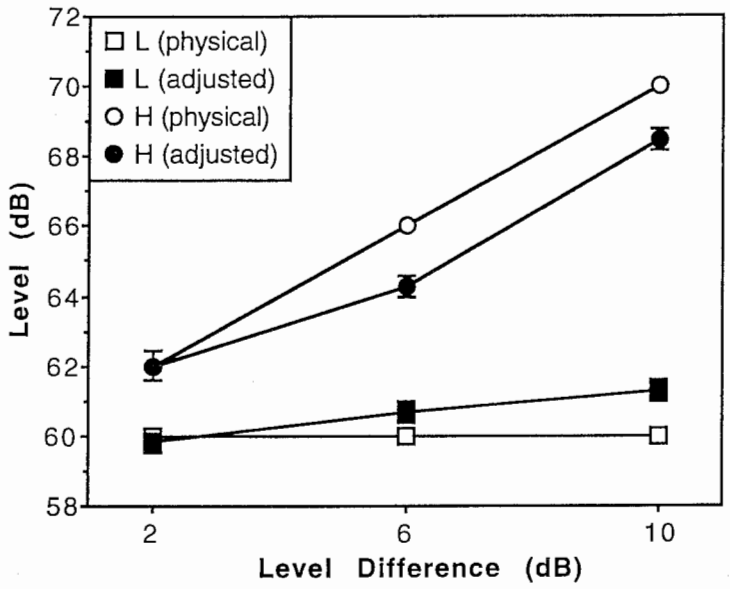


Figure 5. Experiment 4: Physical stimulus levels and mean level adjustments as a function of the level difference between H and L bursts. Vertical bars (where visible) show standard error.

CONCLUSIONS

Experiment 1 demonstrated that a process akin to the "old-plus-new" strategy seems to operate on the alternating level sequences. However, the law by which the loudness is partitioned did not correspond exactly to either pressure or power subtraction. The same law would appear to operate on pure tone and narrow-band noise stimuli and any effects due to the duty cycle of the alternation appear to be attributable to temporal summation of loudness.

Experiment 2 showed that adjustments to single-level continuous and intermittent stimuli are very accurate when test and comparison sequences have the same temporal structure. The level of intermittent sequences were underestimated when compared with a continuous tone, however.

Experiment 3 showed that for dichotically presented alternating sequences in which the continuity illusion is not heard, the level of the softer stream is overestimated by an amount that increases with the level difference between higher and lower levels. No adjustment bias occurs for the higher-level stream. It is important to note that the higher-level stream was adjusted near the physically presented level in the absence of the continuity illusion.

Experiment 4 found that diotic alternating sequences with interspersed silences that break the continuity illusion also resulted in adjustments that were close to the physically presented levels. However the adjustment biases were different in nature from those found for the dichotic sequences. Adjustments for the higher-level stream tended to be low and those for the low-level stream tended to be high. These biases were greater for larger level differences between the streams. The overestimation of the lower-level intermittent stream was not found when this stream was heard as continuous in Experiment 1 under similar conditions. Nonetheless, if we take the bias on the higher-level stream as indicative of a possible bias in adjustments to the intermittent stream in the continuity illusion, we can "correct" those values accordingly. This "correction" moves the data more in the direction of predictions of the power law (being within 1 dB of those predictions for the 6 and 10 dB level differences, though the adjusted level is still about 3 dB too high for the 2 dB difference). The correction moves the data further away from those predicted by the pressure law. It would seem therefore, that the auditory mechanism that subtracts the level of the continuous signal from that of the total level in the H bursts in order to recover the level of the intermittent signal works according to a law that is based on calculations whose units are closer to power than to pressure.

REFERENCES

Bregman, A.S. (1990). *Auditory scene analysis.* (MIT Press, Cambridge, MA).

Houtgast, T. (1972). *J. Acoust. Soc. Am.*, **51**, 1885-1894.

Thurlow, W.R. (1957). *Am. J. Psychol.*, **70**, 653-654.

Thurlow, W.R. & Elfner, L.F. (1959). *J. Acoust. Soc. Am.*, **31**, 1337-1339.

Van Noorden, L.P.A.S. (1975). *Temporal coherence in the perception of tone sequences.* Doctoral dissertation, Eindhoven University of Technology.

Warren, R.M. (1982). *Auditory perception: A new synthesis.* (Pergamon Press, New York).

Warren, R.M. *et al.* (1994). *Perception and Psychophysics*, **55**, 313-322.

Warren, R.M. *et al.* (1972). *Science*, **176**, 1149-1151.

Report 84: file: kyoto.tex, 28 Aug 1994

For the ATR Workshop
A Biological Framework for Speech Perception and Production
16-17 September 1994, Kyoto, Japan

ON THE PERCEPTUAL SEGREGATION OF STEADY-STATE TONES

William Morris Hartmann
Department of Physics
Michigan State University
East Lansing, MI, USA, 48824

ABSTRACT

Human listeners can perceptually segregate two different simultaneous tones, even if the spectra of the two tones are interleaved, and even if the two tones have coincident onsets. This remarkable ability can be profitably studied with psychoacoustical experiments using the mistuned harmonic paradigm, where the listener is required to detect a single mistuned harmonic in an otherwise periodic complex tone. A particular goal of the experiments is to decide whether the segregation process can be best explained as a spectral analysis or a temporal analysis.

Recent mistuned harmonic matching experiments have studied the dependence of segregation on mistuned harmonic number, amount of mistuning, fundamental frequency, tone duration, and level. The data support a temporal model of segregation in which the detection of neural asynchrony plays a preeminent role. The ability to match decreases with increasing mistuned harmonic number in a way that precisely parallels the loss of synchrony observed in physiological recordings from eighth-nerve neurons. Further, mistuned harmonic detection experiments show a nonmonotonic dependence on signal level that resembles the level dependence of multiple synchrony in the eighth nerve.

A question of current interest concerns the tuning of the decision process: whether synchrony/asynchrony is assessed within a narrow frequency channel or whether the process makes comparisons across the entire tonotopic axis. Experimentally, this question has been studied by various manipulations of harmonics that are near neighbors of the mistuned harmonic. Most of the data suggest that the system is tuned, though the channels appear to be broader than typical critical bands.

INTRODUCTION

When we look at two separated objects, images are formed at separated places on the retina where the neural processing of visual stimuli begins. If the angular separation of the objects is great enough the visual system can resolve the separate images, and it registers the fact that there are two objects out there, or anyhow that there are more than one.

When we listen to two distinct sine tones, patterns are formed at separated places on the basilar membrane where the neural processing of auditory stimuli begins. If the frequency separation between the sine tones is great enough (a little less than a critical bandwidth) the auditory system can resolve the separate tones, and it registers the fact that there are two tones.

When we look at two objects that overlap or for which the retinal images are interleaved, the initial neural stimulus is not an adequate basis for identifying, or even counting, the objects. The identification of overlapping objects requires pattern recognition processes with intricate rules of inference (Marr, 1982).

When we listen to two complex tones with interlaced partials the spatial pattern on the basilar membrane (or any other tonotopic coordinate) is not an adequate basis for determining the pitch or tone color of either tone. The pattern recognition process by which individual auditory entities are extracted from a collection of interlaced partials is known as segregation and integration (Hartmann, 1988).

The partials of a periodic complex tone are not normally heard out individually. They fuse perceptually to create a single entity characterized by a pitch and tone color. This is the process of integration. When there are two complex tones, not all the the partials are integrated into one entity. The auditory system appears to sort the partials in a way that creates several different entities. This is the process of segregation.

In listening to speech, music and the sounds of the everyday environment, the most important basis for the segregation of different entities is in the onset. Partials generated by a single source tend to start together, and their starting time is generally slightly different from the starting time of the partials of a different source. It is tempting to suppose that the superb temporal resolution of which the auditory system is capable actually evolved in order to perform just such processes of segregation. The psychophysical effects of asynchronous onsets, especially as they occur in the perception of polyphonic music, have been studied by Rasch (1978, 1979). A related study by McAdams (1989) has shown how common frequency modulation can lead to integration while dissimilar modulation can promote segregation.

A further basis for segregation and integration results from the tendency of many important sounds to be periodic and therefore to have harmonic partials. This is true of the sounds of sustained-tone musical instruments and of the vowels of human speech. The periodicity appears to be a basis for integration that applies to the steady-state part of a sound, independent of onsets.

The significance of periodicity in integrating the harmonics of a single tone and the significance of different periodicities in segregating different tones immediately raises questions about the right kind of model to use in thinking about the process. Should the model

emphasize the periodicity, as it appears in the time domain, or should the model emphasize the harmonic structure in the spectral domain?

Segregation and integration are essentially matters of pitch perception because tones, either one or more, are identified primarily by their pitches. Therefore, it is natural that issues of timing and spectrum that arise in pitch perception also appear in segregation and integration.

The residue theory of pitch perception is a timing model that is specifically designed to deal with the sense of low pitch that one perceives in a periodic waveform. As a model of the fundamental pitches of complex tones, the residue theory is now discredited. However, there are timing models for the pitch of a pure tone which are still useful, especially in explaining the pitches of short tones (Goldstein and Srulovicz, 1977). Timing models of pitch have gained particular credibility recently from studies of pitch perception in individuals with cochlear implants.

A timing model for integration and segregation emphasizes the idea that the various harmonics of a complex tone are phase locked and therefore the neural spikes in different frequency channels are synchronized. It is the synchrony that leads to integration of the channels. Correspondingly, a failure to find an appropriate synchrony in all tuned channels is evidence that leads to the perception of two or more tones. It is evident that segregation and integration can be based upon synchrony or deviations from synchrony only in those tuned channels where the neurons can encode synchrony. This limits the use of synchrony to low frequencies.

The spectral model of segregation and integration is a theory that is consistent with modern approaches to complex tone pitch perception. It focuses on spectral components that are resolved by the auditory system and subjects them to a template fitting process. Different aspects of template fitting are emphasized in pitch models by Goldstein (1973) and Terhardt (1974). The integration/segregation model of Duifhuis, Willems and Sluyter (DWS, 1982) regards the template as a sieve. Spectral components that pass through the sieve are integrated into a complex tone and contribute to the pitch and timbre of the tone. Components that do not pass are segregated into some other percept. In favor of this model is the fact that it is able to segregate two simultaneous speech sounds with the same success as human listeners.

THE MISTUNED HARMONIC EXPERIMENT

To answer questions about timing and/or spectrum as bases for integration and segregation we have studied the detection of mistuned harmonics. Consider a complex periodic tone with a fundamental frequency near 200 Hz and a number of strong harmonics. Such a tone is perceived as a single entity with a low pitch and a bright or buzzy timbre.

If one of the harmonics of the complex tone is slightly mistuned from its correct harmonic frequency, several effects can occur. First, there is a change in pitch of the complex as a whole (Moore *et al.*, 1985). This happens because the auditory system computes the low pitch of a complex tone as a weighted average of frequency information from the

various partials of the tone. Second, there are "beats of mistuned consonances" (Plomp, 1967), reflecting a sensitivity in the auditory system to dynamic phase changes. Most importantly, the mistuned harmonic may be heard out from the complex as a whole as a separate entity. If, for example, the fourth harmonic is mistuned, the listener may become aware of a flute-like tone playing the double octave and accompanying the buzzy tone having the low pitch. It is this latter effect that interest us.

Two psychoacoustical paradigms have been used to study the detection of the mistuned harmonic. One of them is the *mistuned harmonic matching experiment.* Here the listener's task is to match the pitch of the mistuned harmonic in a complex tone by adjusting the frequency of a sine tone. The complex tone and the sine tone are presented successively, not simultaneously. The listener does not know which harmonic is mistuned on a given trial; it might be any one from the fundamental to the 16th harmonic. If the listener matches the mistuned harmonic correctly, he scores a "hit". Otherwise the match is called a "miss." What is particularly attractive about this paradigm is that if the listener can successfully perform a pitch match then we know that he has heard out the mistuned harmonic as a separate entity. It is impossible to make a successful match based upon the other effects of a mistuned harmonic, the shift of the low pitch or the beats of mistuned consonances.

A more efficient paradigm is the *mistuned harmonic discrimination experiment.* Here the listener hears two tones in succession. In one tone the partials are all perfect harmonics. In the other, randomly the first or the second, there is a mistuned harmonic. The listener's task is to identify the tone that includes the mistuned harmonic. Because the task is not specific, one must guard against the artifacts caused by pitch shift cues and beats of mistuned consonances (Moore, *et al.*, 1986). To minimize the role of beats, experiments using the discrimination paradigm use tones with duration less than 100 ms, usually less than 50 ms.

Matching Experiments and Synchrony

Mistuned harmonic matching experiments give evidence that neural synchrony plays an important role in segregating a mistuned partial in a complex tone (Hartmann, *et al.*, 1990). In these experiments the mistuned harmonic number and the fundamental frequency were systematically varied. The data showed that the most important determinant of the listeners ability to segregate is the frequency of the mistuned component. Unlike the predictions of models based upon spectral resolution, and unlike the DWS model, segregation is not a simple function of mistuned harmonic number. Instead, the ability to segregate drops dramatically with increasing frequency, between 2 and 3 kHz. This effect is so dramatic that it even dominates the effect of changing the amount of mistuning, which is the major controlled variable of the experiment. Comparison with synchrony measurements made on auditory neurons in cat shows that the drop in performance on the segregation task precisely parallels the decrease in maximum possible synchrony as a function of frequency. This supports the timing model of segregation, because it is only at frequencies where synchrony is possible that the modulated synchrony, or asynchrony, associated with a mistuned partial can be noticed.

Discrimination Experiments and Autocorrelation

Having identified an important role for synchrony, or neural timing, in the segregation of mistuned harmonics, it is natural to wonder how timing is used in the process. There is information on this question in the results of mistuned harmonic discrimination experiments. These are parametric studies, and there are a lot of parameters to vary: fundamental frequency, mistuned harmonic number, amount of mistuning, duration of the tones, relative phases of the partials, and signal level. As it turns out, all of them matter.

Level effects

Mistuned harmonic detection experiments using mistuned 4th, 5th and 6th harmonics of a 200 Hz complex tone find that the ability to detect a mistuned harmonic is a non-monotonic function of signal level. The function has a maximum at a level of about 40 dB SPL. A possible explanation for this maximum can be found in the neural synchrony studies of Javel (1980) and Greenberg et al. (1983), where it was found that a neuron of the eighth nerve can synchronize to both of two different frequencies if the level is about 40 dB. At higher levels one of the two, usually the tone of lower frequency, dominates the synchrony.

Amount of Mistuning

There is one parameter whose effect seems *a priori* obvious, namely, the amount of mistuning of the mistuned harmonics. When the mistuning is zero there is no basis for discrimination - the two tones are identical. One expects that the greater the mistuning, the more distinguishable the tones should become and the higher the percentages of correct responses in the forced-choice task should be. The experimental test of this idea used a 200 Hz tone with seven harmonics of equal amplitude at an overall level of 40 dB SPL. The duration of the tones was 50 ms. The fourth harmonic was mistuned, and the amount of mistuning was a parameter. As expected, increased mistuning led to improved detection. However, the detectability showed a plateau at a mistuning of 20 Hz, corresponding to one synchrony cycle (20 Hz $\times$ 50 ms = 1). Similarly, when the duration of the tones was reduced to 30 ms, the detection plateau occurred near 33 Hz. Autocorrelator models, such as the tuned autocorrelator model (Hartmann, 1986) or the summary autocorrelogram of Meddis and Hewitt, (1991 a,b) are capable of predicting the plateau at one synchrony cycle.

The Tone Duration

A prediction of the autocorrelator models is that if one does an experiment with the amount of mistuning held constant and the tone duration varied as a parameter, some oscillatory behavior might appear, corresponding to synchrony cycles. Apart from this autocorrelation model, there is no reason to expect anything other than a monotonic improvement in performance with increasing tone duration.

The experiments to test the role of tone duration were identical to the experiments above except that the amount of mistuning of the fourth harmonic was fixed at 20 Hz, and the tone duration was varied. On the basis of the model one expects a plateau in the region of 50 ms, and that is just what the detectability data showed. In some cases, the data actually had a local maximum at the duration of a synchrony cycle.

EVIDENCE FOR TUNED CHANNELS

If it is accepted that mistuned harmonics are segregated from the background on the basis of synchrony anomalies, it remains to discover the locus of those anomalies, whether they occur in tuned channels or at a site that looks across all tuned neural channels. We have studied the question of tuning with two kinds of mistuned harmonic experiments, gap experiments and interference experiments. Both of them are of the discrimination type.

The gap experiments test the idea that a mistuned harmonic is segregated because it fails to exhibit a common synchrony with its neighbors in the same tuned channel. If the neighbors are removed there would no longer be any basis for judging the synchrony within the channel. Gap experiments were done with a mistuned fourth harmonic in which the third harmonic, or the fifth harmonic, or both were missing. The data showed that missing neighbors, especially a missing third harmonic, caused a marked decrease in the detectability of the mistuned fourth.

The interference experiments attempt to interfere with the detection of a synchrony anomaly, caused by a mistuned target harmonic, by some other mistuning, which also leads to a synchrony anomaly. Specifically, the task was to detect a mistuned sixth harmonic, mistuned by a small amount on one of the two intervals, in the presence of a mistuned second harmonic, mistuned by a large amount on both intervals. The data showed a small decrement in detectability of the mistuned fourth caused by the interference of the mistuned second. Considerably greater decrement occurred when the interfering mistuned harmonic was changed to the sixth. By contrast, interference from a mistuned eighth was negligible. Thus, interference exhibits tuning, suggesting that synchrony is evaluated in tuned channels, indicating auditory filtering. Both the gap experiments and the interference experiments suggest a filter with a high frequency slope that is considerably steeper than the low frequency slope.

CONCLUDING DISCUSSION

The perceptual operations of integration and segregation can be approached in terms of both tonotopic spectral template fitting models and neural synchrony timing models. The mistuned harmonic experiment gives considerable evidence to support the idea that listeners segregate mistuned harmonics on the basis of anomalies in neural synchrony. The rapid decline in the ability to detect a mistuned harmonic with increasing frequency parallels the loss of synchrony in the mammalian ear as observed in physiological studies. The nonmonotonic dependence of detectability on signal level parallels the dependence of multiple synchronies observed physiologically. Structure in the dependence of detectability upon the amount of mistuning and the duration of the stimulus can both be understood from an autocorrelation model for synchronous neural spikes. By contrast, the tonotopic template fitting models do not account for the above effects. To the extent that the mistuned harmonic experiment is a representative segregation operation, neural synchrony is an important part of the integration and segregation of the complex tones in music and speech.

Mistuned harmonic detection experiments employing spectral gaps or interfering tones suggest that neural synchrony is evaluated in tuned channels. This means that synchrony anomalies caused by a mistuned harmonic are recognized by comparison with only a small set of neighboring harmonics. The experiments do not suggest that there is an overseer that scans across the entire tonotopic axis. Ultimately, of course, there must be some process that combines the information from different tuned channels to form the integrated entities that we recognize as music and speech. The mistuned harmonic experiments, however, indicate that there is preliminary synchrony evaluator that is tuned. On the other hand, a long series of experiments on the pitches of mistuned harmonics, none of which was discussed in this paper, appear to be suggesting that mistuned harmonics are actually perceived with respect to the entire background of the other harmonics in the complex. In fact, the most straightforward explanation for the recent pitch results lies in a model that resembles a template fitting model (Lin and Hartmann, 1994). The resolution of the detection data with the recent pitch data must await another time, another workshop, and another paper.

ACKNOWLEDGEMENT

This work was supported by the United States National Institutes of Health.

REFERENCES

Duifhuis, H., Willems, L.F., and Sluyter,R.J. (1982) "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," J. Acoust. Soc. Am., **71**, 1568-1580.

Goldstein, J.L. (1973) "An optimal processor theory for the central formation of the pitch of complex tones," J. Acoust. Soc. Am. 54, 1496-1516.

Goldstein, J.L. and Srulovicz, P. (1977) "Auditory-nerve spike intervals as an adequate basis for aural spectrum analysis," in *Psychophysics and Physiology of Hearing*, ed. E.F. Evans and J.P. Wilson, (Academic, New York) pp.337-345.

Greenberg, S., Geisler, C. D., and Deng, L. (1983) "Filter characteristics of low-frequency cochlear nerve fibers as determined by synchrony response patterns to two-component signals." J. Acoust. Soc. Am. (abst) 74, S7.

Hartmann, W.M. (1986) "Pitch and the perceptual organization of complex spectra," J. Acoust. Soc. Am. (abst.) **79**, S65.

Hartmann, W.M. (1988) "Pitch perception and the segregation and integration of auditory entities," in *Auditory Function*, ed. G.M. Edelman, W.E. Gall and W.M. Cowan (Wiley, New York) pp.623-645.

Hartmann, W.M., McAdams, S., and Smith, B.K. (1990) "Matching the pitch of a mistuned harmonic in an otherwise periodic complex tone," J. Acoust. Soc. Am. **88**, 1712-1724.

Javel, E. (1980) "Coding of AM tones in the chinchilla auditory nerve: Implications for the pitch of complex tones," J. Acoust. Soc. Am. **68**, 133-146.

Johnson, D.H. (1980) "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," J. Acoust. Soc. Am. **68**, 1115-1122.

Lin, Jian-Yu and Hartmann, W.M. (1994) "The pitches of mistuned harmonics," J. Acoust. Soc. Am. (abst) **95**, pp 2965.

Marr, D. (1982) *Vision*, W.H. Freeman, San Francisco.

McAdams, S. (1989) "Segregation of concurrent sounds. I: Effects of frequency modulation coherence." J. Acoust. Soc. Am. **86**, 2148-2159.

Meddis, R. and Hewitt, M.J. (1991a) "Virtual pitch and phase sensitivity of a computer model of the auditory periphery I." J. Acoust. Soc. Am. **89**, 2866-2882.

Meddis, R. and Hewitt, M.J. (1991b) "Virtual pitch and phase sensitivity of a computer model of the auditory periphery II." J. Acoust. Soc. Am. **89**, 2883-2894.

Moore, B.C.J., Glasberg, B.R. and Peters, R.W. (1985) "Relative dominance of individual partials in determining the pitch of complex tones," J. Acoust. Soc. Am. **77**, 1853-1859.

Moore, B.C.J., Peters, R.W., and Glasberg, B.R. (1986) "Thresholds for hearing mistuned partials as separate tones in harmonic complexes," J. Acoust. Soc. Am. **80**, 479-483.

Plomp, R. (1967) "Beats of mistuned consonances," J. Acoust. Soc. Am. **42**, 462-474.

Rasch, R.A. (1978) "The perception of simultaneous notes as in polyphonic music," Acustica **40**, 21-33.

Rasch, R.A. (1979) "Synchronization in performed ensemble music," Acustica **43**, 121-131.

Terhardt, E. (1974) "Pitch, consonance and harmony," J. Acoust. Soc. Am. **55**, 1061-1069.

ON THE PERCEPTUAL DISTANCE BETWEEN SPEECH SEGMENTS

Oded Ghitza and M. Mohan Sondhi

AT&T Bell Laboratories
Murray Hill, New Jersey 07974, USA

ABSTRACT

For many tasks in speech signal processing it is of interest to develop an objective measure that correlates well with the perceptual distance between speech segments. (By speech segments we mean pieces of a speech signal, of duration 50-200 milliseconds. For concreteness we will consider a segment to mean a diphone.) Such a distance metric would be useful for low bit rate speech coders because perturbations introduced by such coders typically last for several tens of milliseconds. It would also be useful for automatic speech recognition on the assumption that mimicking human behavior will improve recognition performance. Yet a third use for such a metric would be to define a just noticeable difference for diphones (a "phonemic" JND). (If a diphone is perturbed, how far from the original must the perturbed diphone be, in order to be perceived as a different diphone?) In this talk we will describe our attempts at defining such a metric.

I. INTRODUCTION

This paper is concerned with psychoacoustical experiments relevant to the perception of speech. In the past, such experiments have been concerned with the perception of what we may call "frame level" properties, that is, properties that can be derived by examining speech through a short (20-30 millisecond) time window. Typically these experiments are concerned with (a) masking of steady state signals by other steady state signals (e.g., masking of tones by noise, noise by a tone, etc.); or (b) measurement of the just noticeable difference (JND) of some steady state property (e.g., JND for amplitude or pitch of a tone, JND for formant frequencies, etc.). Speech, however, is a highly nonstationary signal, and it is not at all clear how masking properties and JND's change due to this nonstationarity. Therefore these studies are of limited application to problems such as speech coding at low bit rates and automatic speech recognition. Almost all progress in these areas has come from application of signal processing techniques, with little help from psychophysics.

In this paper we will describe our attempts at improving this situation. We will consider psychophysical experiments involving "segment level" properties, where by segments we mean pieces of speech signals with durations of the order of 50-200 milliseconds. For concreteness we will consider diphones, although longer segments could be studied by similar methods. The main problem we will address here is the derivation of a "perceptual distance" between two such segments of (in general) unequal duration. Useful measures of distance between two speech signals have, of course, been proposed in the past. Our point of view is different, however, in that we would like the distance to have perceptual relevance.

A measure of the perceptual distance would be of interest in its own right. It would also have practical applications. For instance, perturbations introduced by low bit rate speech coders extend over segment length intervals. The design and evaluation of such coders should therefore benefit from the derivation of a perceptual distance of the type considered here. Also, we believe that a

perceptual distance would provide a robust measure for automatic speech recognition.

Our approach to the problem may be described briefly as follows: The paradigm used is the Diagnostic Rhyme Test (DRT). The word pairs in the DRT are modified by interchanging judiciously selected time-frequency regions (tiles). This modified database is used in the standard DRT, and the error patterns induced by these changes are recorded. The same DRT is then *simulated* by an array of speech recognizers based on a parametric distance function. The parameters are optimized so as to mimic the error patterns of the human subjects. In the following sections we will describe these steps in somewhat greater detail.

In Section II we will summarize the DRT paradigm, which is well known as a tool for the evaluation of speech coders. We will also describe the way in which we simulate the paradigm by replacing the human subjects by an array of automatic speech recognizers. In Section III we will describe the interchange of time-frequency tiles alluded to above. This "tiling" experiment has also been described in a recent paper, so we will only summarize it briefly. Finally, in Section IV, we will discuss the optimization procedure and the degree of success achieved by the simulation in mimicking human error patterns.

II. THE DIAGNOSTIC RHYME TEST (DRT)

Psychophysics

For the psychophysical paradigm we have chosen the DRT, which was first suggested by Voiers [8], and which has been in extensive use for evaluating speech coders. We will discuss our reasons for this choice after first describing the test.

In the DRT, Voiers uses 96 pairs of confusable words spoken by several male and female speakers. All the words are of the CVC type, and the words in each pair differ only in the initial consonant. [More recently, Voiers has assembled another database in which the words in each pair differ in the *final* consonant. The corresponding test based on this database is termed the Diagnostic ALliteration Test (DALT). Yet a third database has been recently developed in which the words are of the VCV type, and the words in each pair differ only in the *medial* consonant. The corresponding test is termed the Diagnostic Medial Consonant Test (DMCT). In our experiments we have used the DRT and DALT, but have not yet utilized the DMCT. In this paper, to avoid repetitious statements, we will describe the experiments in terms of the DRT. All statements apply, with obvious modification, to the DALT and DMCT as well.]

The target diphones (initial for DRT, final for DALT and medial for DMCT) are equally distributed among six phonemic distinctive features (16 word pairs per feature) and among eight vowels. The feature classification follows the binary system suggested by Jakobson, Fant and Halle [5]. The dimensions are voicing, nasality, sustension, sibilation, graveness and compactness, and the target consonants in each pair differ in the presence or absence of one of these dimensions. An explanation of these attributes, as well as the complete list of words for the DRT and DALT may be found in [2].

The database is used in a very carefully controlled psychophysical procedure. The listeners are well trained and quite familiar with the database, including the voice quality of the individual speakers. A one interval two alternative forced choice paradigm is used. A word pair is selected at random and displayed as text on a screen. One of the words in the pair (selected at random) is next presented aurally, and the subject is required to indicate which of the two words was heard. The procedure is repeated until all the words in the database have been presented. The errors

made by the subjects are recorded and may be analyzed in various ways, as discussed in Section IV.

The conditions of the paradigm are such that the subject is given as much of the "cognitive" information about the stimuli as possible, so that the errors may be attributed entirely to the peripheral processing in the auditory pathway. This is the aspect of perception that we are interested in. And the fact that the DRT allows us to focus on it, is the main reason for our choice of this paradigm.

Simulation

As mentioned above, the subject is given as much of the cognitive information about the stimuli as possible. We make the assumption that the subject is able to utilize this information. Thus, when presented with an utterance to be identified, the subject is able to process it through two models (one for each word in the pair displayed visually) and choose the one judged closest. To simulate the DRT, therefore, we implement an array of automatic speech recognizers, one for each pair of words in the database. The unknown utterance is examined by the appropriate recognizer and scored by the models for each of its two words. The utterance is classified as the word whose model gives the best score.

This method of simulating the DRT has been described in [3], and the reader is referred to that article for details. The particular type of speech recognizer that we use in the simulation is also described in a recent article [4], so we will not describe it in detail here. Suffice it to mention that the recognizer comprises Hidden Markov Models with nonstationary states, where each state is a template of a diphone. When used in the DRT, the recognizer is essentially a recognizer of the initial diphone, since the second diphone of the CVC is identical for the two words in each pair. Thus correct recognition occurs if and only if the initial diphone of the utterance is closer to the model for the initial diphone of the correct word than to the model of the other word of the pair.

The errors made in this simulation are entirely governed by the definition of distance between the test diphone and the model diphone. This distance may be defined in parametric form in a variety of ways. Our definition is as follows: Define a diphone as a sequence of feature vectors — one for each frame. Our choice of feature vector is a 30-dimensional EIH with ERB-rate bin allocation [1]. Let $\mathbf{X} \equiv [\mathbf{x}_1 , \mathbf{x}_2 , \cdots , \mathbf{x}_N]$ and $\mathbf{S} \equiv [\mathbf{s}_1 , \mathbf{s}_2 , \cdots , \mathbf{s}_M]$ be the sequences of feature vectors for the test utterance and the template (or state), respectively. For the template we define two matrices: M_c for the consonant, and M_v for the vowel. In terms of M_c and M_v we define the distance between two vectors $\mathbf{x}$ and $\mathbf{s}$ as

$$d(\mathbf{x} , \mathbf{s}) = (\mathbf{x}-\mathbf{s})^t M^t M(\mathbf{x}-\mathbf{s})$$

where M is replaced by M_c if $\mathbf{s}$ is in the consonant portion of the template, and by M_v if it is in the vowel portion. With this definition of distance between vectors, the distance between the sequences $\mathbf{X}$ and $\mathbf{S}$ is defined as

$$D(\mathbf{X},\mathbf{S}) = \min_{n(m)} \sum_{1}^{M} d(\mathbf{x}_{n(m)} , \mathbf{s}_m).$$

This is the usual Dynamic Time Warp (DTW) distance, except for the introduction of the matrices M.

The matrices M_c and M_v may, in general, be different for every template. This, however, is not feasible. Note that the total number of diphones is on the order of 2000 in English. Even the number of diphones in the DRT and DALT is 192. We therefore restrict the number of matrices by using the same sets for diphones with "similar" properties. At present we group together

consonants into six categories (bilabial, labio-dental, dental, alveolar, palatal and velar) and the vowels into four categories (low back, high back, low front amd high front). This gives us 24 classes of diphones, and we assign a set of matrices to each such class.

We also have the freedom of choosing the structure of the matrices M_c and M_v. Again, it is not feasible to use full 30×30 matrices. We have tried diagonal and tridiagonal structures for them.

When all the parameters in all the matrices have been specified, the definition of D gives us a parametrized distance which depends on the template (or state). The choice of parameters is optimized so as to match the error patterns of the simulated DRT to the error patterns of the human subjects, in the experiments to be defined in the next section.

III. THE TILING EXPERIMENT

The psychophysical experiment used in our search for the perceptual distance is what we call the "tiling" experiment. Details of this experiment may be found in [2]. Briefly, we divide the time-frequency plane into non-overlapping regions called "tiles" that cover the target diphone in each pair of words in the DRT (or DALT). Ideally, one should use many small tiles, but the experiments become increasingly time consuming and expensive with increasing number of tiles. From considerations of feasibility, we decided that we could use six tiles. The six regions were chosen as illustrated in Fig. 1. The selection was made on the basis of the following rough reasoning: On the time axis a break at the boundary between the C and V is an obvious choice. The break at 1 kHz is suggested by the known change in the properties of nerve firings at approximately this frequency. The break at 2.5 kHz corresponds roughly to the upper limit of the second formant frequency [7].

We interchange a tile (or some combination of tiles) between the target diphones of each pair in the database, as illustrated in Fig. 2. A total of 14 different distorted versions of the database are created in this way. Each of these versions is used in both the psychophysical and the simulated DRT, as described in Section II. The error pattern induced by each of these distortions is recorded. Some examples of the patterns of errors along the Jakobson, Fant, Halle dimensions are shown in Fig. 3.

IV. THE OPTIMIZATION

Let us denote by **M** the set of all the parameters entering the 24 sets of matrices M_c and M_v defined in Section II. Starting with a trial set of parameters, the error patterns for the simulated DRT are computed for each of the distorted databases of the tiling experiment described in the last section. The parameters are optimized so as to minimize the difference between human and machine performance. This difference is measured by a cost function , C, defined as the squared difference between the human and machine errors, accumulated over all 14 DRTs with the 14 tiled versions. Thus

$$C = \sum_t \sum_i (h_{ti} - m_{ti})^2,$$

where h_{ti} and m_{ti} are the errors made by the human and machine respectively, for the i-th dimension and the t-th tiling. In order to make C a continuous function of the parameters **M**, a "soft" definition of error is used. Thus if a test diphone is at a distance D_1 from the correct model, and at D_2 from the incorrect model, then the soft error e_s is defined as

$$e_s = \frac{1 + arctan\ [\ k\ (D_1 - D_2)]}{2}$$

As k becomes large e_s approaches 0 if the test is closer to the correct diphone, and 1 if it is closer to the incorrect diphone.

Since it is not possible to analytically compute the gradient of the cost function C, we use a gradient-less optimization procedure. The one we have chosen is the simplex method [6].

REFERENCES

[1] Ghitza, O. (1994). "Auditory models and human performance in tasks related to speech recognition and speech coding", *IEEE trans. on Speech and Audio, SAP-2(1).* Special issue on Neural networks for Speech Processing, 115-132.

[2] Ghitza, O. (1993). "Processing of spoken CVCs in the auditory periphery: I. Psychophysics", *Journal of the Acoustical Society of America, 94(5)*, 2507-2516.

[3] Ghitza, O. (1993). "Adequacy of auditory models to predict internal human representation of speech sounds", *Journal of the Acoustical Society of America, 93(4)*, 2160-2171.

[4] Ghitza, O. and Sondhi, M. M. (1993). "Hidden Markov Models with Templates as Nonstationary States: An Application to Speech Recognition", *Computer Speech and Language, 7(2)*, 101-119.

[5] Jakobson, R., Fant, C. G. M. and Halle, M. (1952). "Preliminaries to speech analysis: the distinctive features and their correlates", *Technical Report No. 13, Acoustic Laboratory*, Massachusetts Institute of Technology, Cambridge, Mass.

[6] Nelder, J. A. and Mead, R. (1965). "A Simplex Method for Function Minimization", *Computer Journal, 7*, 308-313.

[7] Peterson, G. E. and Barney, H. L. (1952). "Control methods used in a study of the vowels", *Journal of the Acoustical Society of America, 24*, 175-184.

[8] Voiers, W. D. (1983). "Evaluating processed speech using the Diagnostic Rhyme Test", *Speech Technology, 1(4)*, 30-39.

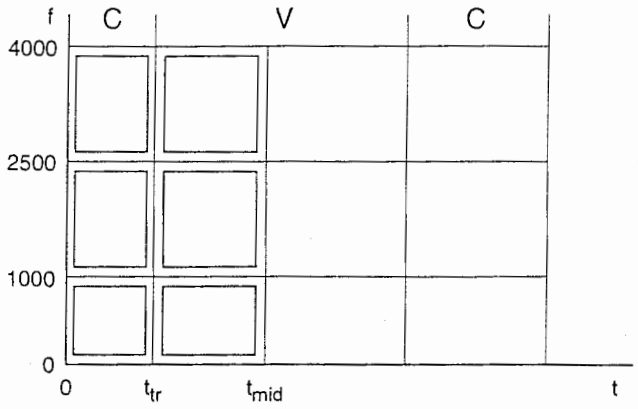


Figure 1. A diagram of the time-frequency region occupied by a spoken CVC word. The time-frequency region of the initial diphone is sub divided into 6 "tiles". The frequency boundaries (from the bottom up) are 0 Hz, 1000 Hz, 2500 Hz and the highest frequency in the band, say 4000 Hz. The time landmarks are (from left to right) the beginning of the word (t = 0), the transition from the initial consonant to the vowel ($t = t_{tr}$) and the mid-point of the vowel ($t = t_{mid}$). For stop consonants, t_{tr} is the transition from the stop release to the vowel.

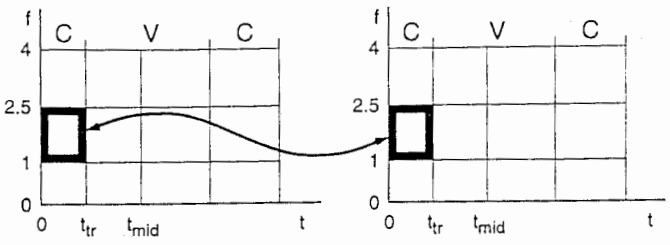


Figure 2. A diagram of the time-frequency region occupied by a prototype DRT word-pair, where the regions corresponding to the initial diphones are divided into 6 tiles each. The interchange of one of the tiles is illustrated.

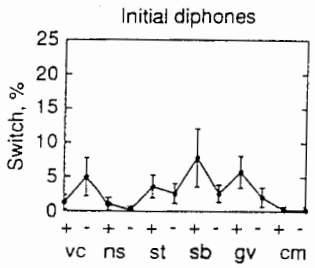


Figure 3(a). The average and the 95% confidence interval for the DRT without any interchange of tiles. The abscissa of every plot indicates the 12 phonemic categories: "vc" is for Voicing, "ns" for Nasality, "st" for Sustention, "sb" for Sibilation, "gv" for Graveness and "cm" for Compactness". The "+" sign stands for attribute present and the "-" sign for attribute absent. The ordinate is termed "switch", and it represents the number of words in the category that, when played to the listener, were judged to be the opposite word in the word pair (i.e., the listener "switched" to the opposite category). The switch is represented as a percentage of 16 (the total number of words per phonemic category).

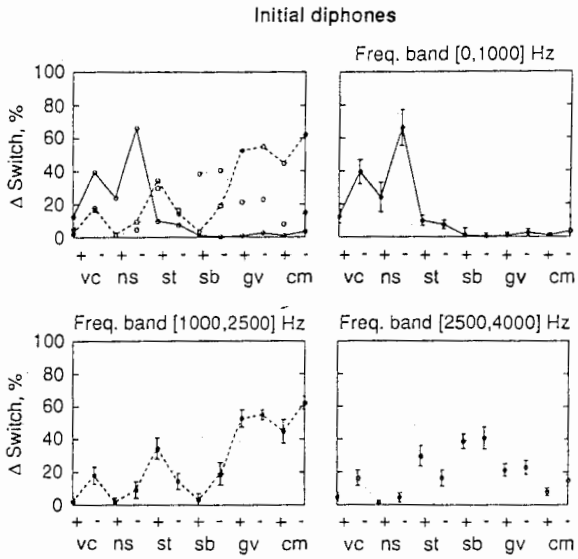


Figure 3(b). Human performance on the DRT database, under interchange of each frequency band over the entire diphone. The upper left plot is a summary of the other 3 plots, with the confidence-interval bars omitted. The abscissa is as in Fig. 3(a). The ordinate is termed "Δ*switch*", since it represents the *additional* number of switches, relative to the baseline version, that occured due to the particular interchange operation. Note that the line connecting the measurements is only for display purposes, to enable the reader to distinguish between error patterns that belong to a particular interchange condition. The upper right plot shows the amount of Δ*switch*, in percent, under interchange of the 0-1kHz band. The lower left plot is for the 1 kHz - 2.5 kHz band, and the lower right plot is for the 2.5 kHz - 4 kHz band. Notice that *Voicing* and *Nasality* are strongly correlated with the first frequency band of the diphone, *Graveness* and *Compactness* with the second frequency band of the diphone, and *Sibilation* with the third frequency band of the diphone.

Neuronal basis for temporal and spectral pitch integration observed in the primary auditory cortex of the Japanese monkey

Hiroshi Riquimaroux
Neural Systems Laboratory, Frontier Research Program, The Institute of Physical and Chemical Research (RIKEN), Wako, Saitama 351-01, Japan.
e-mail: rikimaru@murasaki.riken.go.jp

The temporal structure of speech sounds appears to be very important for speech perception; *e.g.*, the voice onset time, communications through telephone, communications with the cochlear implant. In this workshop, quite a few psychoacousticians have well demonstrated that the temporal structure of sounds is essential for our auditory sensation, typically for pitch extraction. So, I will discuss about the pitch extraction from a neurophysiological point of view and show you neurophysiological correlates to indicate how and where an integration of temporal and spectral information takes place in the central auditory system. The perception of the missing fundamental (f0) generated by combined successive higher harmonics is a well known example for pitch sensation created by temporal information, *temporal pitch* rather than spectral information, *spectral pitch* (Seebeck, 1841; Schouten, 1938; Fig. 1). Although the phase-locked neural firings corresponding to the temporal pitch exist in the temporal discharge pattern of the cochlear nerve fibers (Delgutte, 1980), it has been believed that the missing f0 is created not by the peripheral but by the central auditory system (Moore, 1989). The fact that a low f0 and successive higher harmonics without the f0 create the same pitch sensation implies that we may have a neuron in the central auditory system which is responsible for both spectrally and temporally created pitches. In other words, the temporal pitch and the spectral pitch are likely to be co-place-coded in the central auditory system. Indeed, human lesion studies have suggested that the primary auditory cortex (AI) may play an important role in the perception of the missing f0. Patients with an impaired AI area have difficulties in perceiving the missing f0. However, those who have a lesion in the temporal lobe but an intact AI area do not show problems in perceiving the missing f0 (Zatorre, 1988; Bharcha *et al.*, 1993). More, a recent auditory-evoked magnetic field study has indicated that both a combination of successive higher harmonics without the f0 and the f0 itself are processed by the same area in the auditory cortex of humans (Pantev *et al.*, 1989). However, neuronal or cellular evidences have not been shown, previously.

The Japanese monkey is reported to have a similar missing f0 sensation as humans (Tomlinson and Schwarz, 1988). Recent studies of our laboratory have demonstrated that the neuron in AI of the Japanese monkey, whose best frequency is the f0 ($\leq$ 500 Hz), appears to be also sensitive to the missing f0 generated by

successive higher harmonics (Fig. 2). However, the AI neuron does not respond to these higher harmonics themselves when they are presented alone (Fig. 2). In other words, the spectral pitch and the temporal pitch are evidently co-place-coded in AI. (Riquimaroux and Hashikawa, 1994). The synthesized sound made of the successive higher harmonics (H condition) has a periodicity in the amplitude envelope corresponding to the f0 (Fig. 1B). Thus, the periodicity of the sound amplitude might have an important role for the temporal pitch. However, despite having the identical periodicity in the amplitude envelope, the sound made of successive higher harmonics of the same f0 but of very high frequencies would little excite the same AI neuron (Fig. 2B). So, the periodicity in the amplitude envelope may not be the only parameter that controls the temporally created pitch. Also, the higher harmonics appear to have to be within a certain frequency range to generate the same pitch as the f0. This limitation for the absolute frequency of the higher harmonics is similar to the tendency observed in previous human psychoacoustical studies (Zwicker and Fastl, 1990; Fig. 3). Further, the AI neuron is much less sensitive to a combination of shifted successive higher harmonics (SH condition) with the same spacing frequency as the missing fundamental case (Figs. 4b, 5B and 6). Actually, the periodicity of the amplitude envelope of the SH condition is the same as the H condition. Different from the H condition, we psychoacoustically do not perceive the f0 pitch with the SH condition. Thus again, the periodicity of the amplitude envelope may not be the only essential component to produce the temporal pitch although periodicity sensitive neurons have been found in the inferior colliculus (Langner, 1988). It is hard to conclude only from these data but the autocorrelation in temporal structure might play a role for the temporal pitch extraction. The data imply that the temporal pitch is generated in the central auditory system not in the cochlea and integrated with the spectral pitch to the exent shown here above the inferior colliculus and at or below the AI. More, the missing f0 pitch can be created by a dichotic presentation (Houtsma and Goldstein, 1972). So, it would be of great interest to see how AI neurons behave when odd higher harmonics are presented to one ear and even higher harmonics are given to the other ear (Fig. 5E). How the H condition with a low-pass masker (Fig. 5D) can create an intact missing f0 pitch is still unknown.

acknowledgement

The research was supported by Frontier Research Program, RIKEN and The Sound Technology Promotion Foundation.

Reference

Bharcha, J. J., Tramo, M. J. and Zatorre, R. J. (1993) Abstraction of the missing fundamental following bilateral lesions of auditory cortex. *Soc. Neurosci. Abstr.* **19**: 1687.

Delgutte, B. (1980) Representation of speech-like sound in the discharge pattern of auditory-nerve fibers. *J. Acoust. Soc. Am.* **68**: 843-857.

Houtsma, a. J. M. and Goldstein, J. L. (1972) The central origin of the pitch of complex tones: Evidence from musical interval recognition. *J. Acoust. Soc. Am.* **51**: 520-529.

Langner, G. (1988) Physiological properties of units in the cochlear nucleus are adequate for a model of periodicity analysis in the auditory midbrain. in: *Auditory Pathway*, Syka, J. and Masterton, R. B. (eds.), Plenum, New York, pp207-212.

Moore, B. C. J. (1989) *An Introduction to the Psychology of Hearing.* Academic Press, London, pp 158-193.

Pantev, C., Hoke, M., Lütkenhöner, B. and Lehnertz, K. (1989) Tonotopic organization of the auditory cortex: Pitch versus frequency representation. *Science* **246**: 486-488.

Riquimaroux, H. and Hashikawa, T. (1994) Units in the primary auditory cortex of the Japanese monkey can demonstrate a conversion of temporal and place pitch in the central auditory system. *J. Phys. IV* C5 **4**: 419-425.

Schouten, J. F. (1938) The perception of subjective tones. *K. ned. akad. Wet. Proc.* **41**: 1086-1093.

Seebeck, A. (1841) Beobachtungen über einige bedingungen der entstehung von tönen. *Ann. Phys. Chem.* **53**: 417-436.

Tomlinson, R. W. W. and Schwarz, D. W. F. (1988) Perception of the missing fundamental in nonhuman primates. *J. Acoust. Soc. Am.* **84**: 560-565.

Zatorre. R. J. (1988) Pitch perception of complex tones and human temporal-lobe function. *J. Acoust. Soc .Am.* **84**: 566-572.

Zwicker, E. and Fastl, H. (1990) *Psychoacoustics.* Springer-Verlag, Berlin.

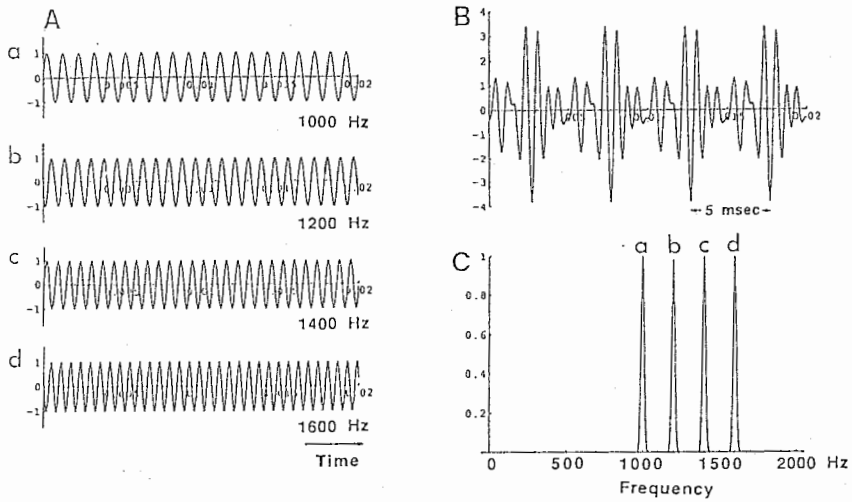


Fig. 1 A combination of four higher harmonics to produce the "missing fundamental". **A:** 1000, 1200, 1400 and 1600 Hz sinusoidal signals (pure tones) are illustrated in a, b, c and d, respectively. **B:** Temporal wave pattern produced by adding 1000, 1200, 1400 and 1600 Hz. A periodicity of 5 ms, corresponding to 200 Hz, is visible. **C:** Power spectrum of the synthesized wave shown in B. Peaks for 1000, 1200, 1400 and 1600 Hz can be observed but no peak for the missing fundamental (200 Hz , broken line).

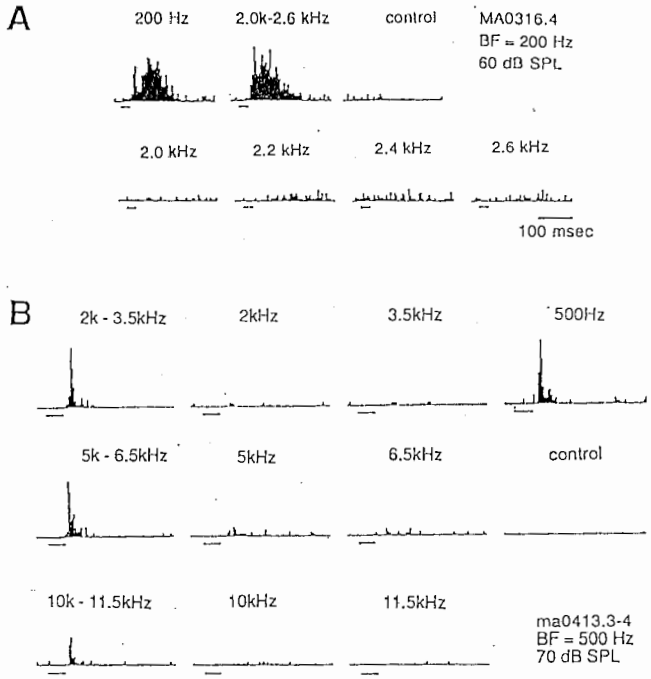


Fig. 2 Responses to the fundamental frequency (fo), combinations of higher harmonics without fo and each higher harmonic component. **A, top row:** from left responses to 200 Hz, a complex of 2.0 + 2.2 + 2.4 + 2.6kHz and control. **bottom row:** responses to 2.0, 2.2, 2.4 and 2.6 kHz. BF = 200 Hz, Intensity = 60 dB SPL. **B, top row:** responses to a complex of 2.0 + 2.5 + 3.0 + 3.5 kHz, 2.0 kHz, 3.5 kHz and 500 Hz. **middle row:** responses to a complex of 5.0 + 5.5+ 6.0 + 6.5kHz, 5.0 kHz, 6.5 kHz and control condition. **bottom row:** responses to a complex of 10.0 + 10.5 + 11.0 + 11.5kHz, 10.0 kHz and 11.5 kHz. BF = 500 Hz. Intensity = 70 dB SPL.

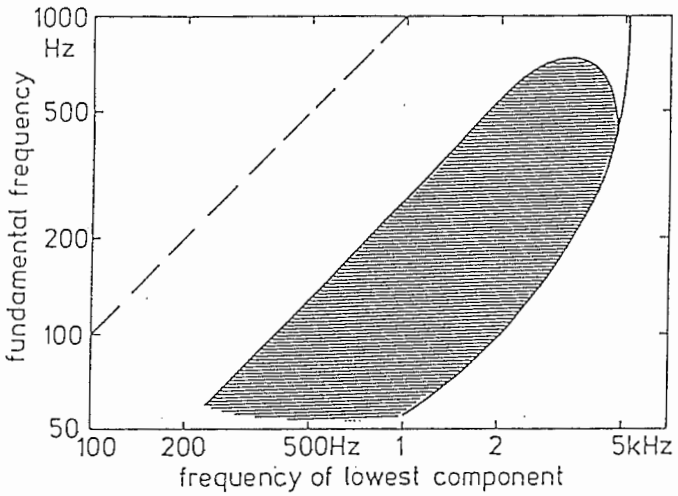


Fig. 3 Region where the missing fundamental pitch exists. Fundamental frequency as a function of the lowest component in the higher harmonic complex. Shaded area: region where the missing fundamental exists. Broken line: fundamental frequency. (from Zwicker and Fastl, 1990)

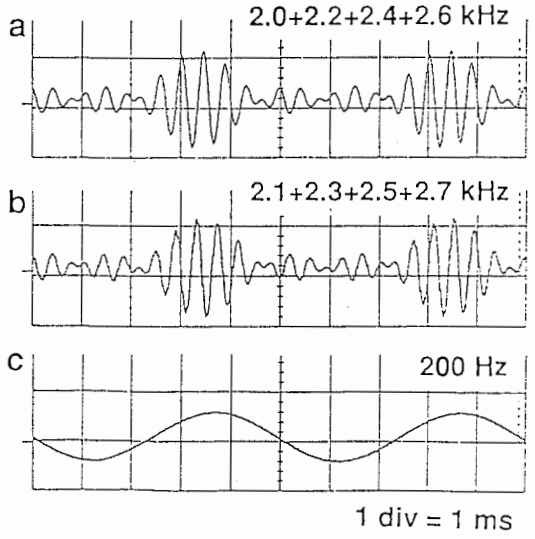


Fig. 4 Temporal wave patterns. **a:** a combination of 2.0, 2.2, 2.4 and 2.6 kHz, which has a periodicity of 200 Hz, which creates the pitch of 200 Hz. **b:** a combination of 2.1, 2.3, 2.5 and 2.7 kHz, which has an envelope with a periodicity of 200 Hz but does not generate 200 Hz pitch. **c:** 200 Hz tone.

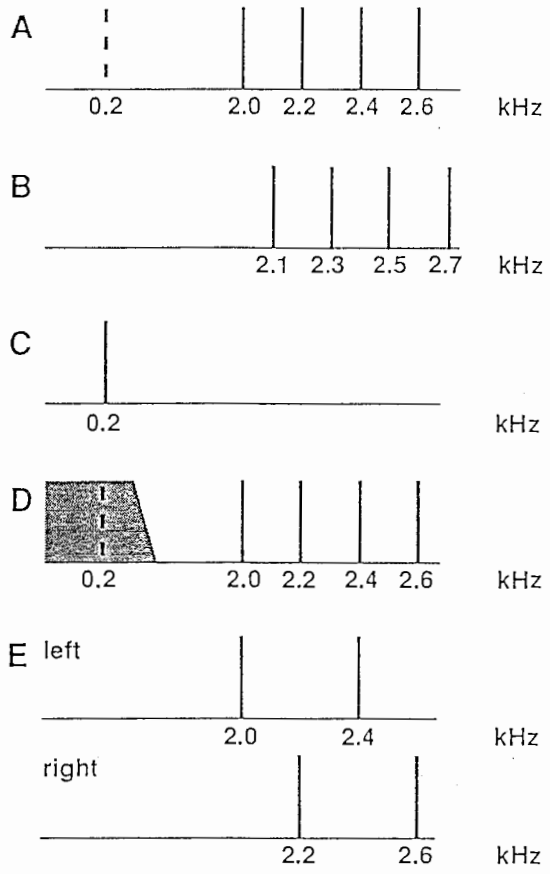


Fig. 5 Schematic power spectrum of synthesized waves. **A, B** and **C** correspond to a, b and c in Fig. 4. **D:** Low-pass noise is added to the wave in A. **E:** Dichotic presentation of higher harmonics of 200 Hz. Odd harmonics to the right while even ones to the left. The missing fundamental is shown in a dotted line.

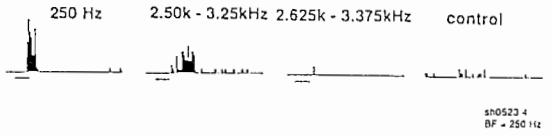


Fig. 6 A comparison between responses to a single tone (250 Hz, the best frequency), a harmonic condition (2.5 + 2.75 + 3.0 + 3.25 kHz) and an inharmonic condition (2.625 + 2.875 + 3.125 + 3.375 kHz). The neuron does not well respond to the inharmonic combination.

Comments on three considerable questions in a biological framework for speech perception

Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology
15 Asahidai, Tatsunokuchi, Ishikawa 923-12, Japan

1. What are considerable questions?

At the end of the first day of the workshop, Dr. Ghitza concluded the first day discussions and presented the following considerable questions to construct an efficient relationship between psychophysics and speech processing, such as speech recognition and coding.

1) What objective criteria can represent the merit of auditory models?
2) What method or model can treat an entity of 200 ms or more?
3) How do we deal with "speech-pitch", rather than dealing with "music-pitch"?

I am trying to offer some comments on questions (1) and (2) above and to provide an article that describes the relations between temporal fine structures and sound segregation techniques from the engineering viewpoint.

2. What objective criteria can represent the merit of auditory models?

Although, recently, the merit of auditory models is being measured by the use of HMM or DTW in the speech recognition stage, what do the people who are using such kind of criteria measure? If these auditory models are to be used for the front-end of a speech recognizer and the merit of the auditory models is evaluated only for HMM and DTW, this measure may be sufficient for objective criteria. However, if we want to use these models for the front-end of a speech coder, who will guarantee its performance?

I propose that we investigate which physical values are emphasized by using auditory models and whether the physical values are useful or not for each application. Prof. Stern showed us that a cross correlation modeling human binaural perception works well to increase the signal-to-noise ratio (SNR) and an increase of SNR is useful for recognizing unclean speech. This is a typical case.

These solutions might be incorrect for psychologists or physiologists. Their primary purpose is to model auditory systems as faithfully as possible, and is not to extract useful values for applications. Thus, they integrate non-linear stages into the models. Prof. Meddis introduced some interesting physiological models of auditory periphery. These models are faithful models for auditory physiology and have some non-linearity. I agree that faithful models are necessary and useful for psychological/physiological research. We have to consider, however, that significant articles for psychology and physiology are not often equally significant for engineering.

Let us focus on speech coding as an application. It is significant, in this application, to extract physical values which represent speech characteristics well and which are handled easily. Dr. Slany showed us some simulated results of reconstructed speech waveforms from cochleagrams. This result is an example well resolved. However, in general, the more the non-linearity increases, the more difficult the reconstruction is. If physical values through complex models are very useful, we shall use these models for coding and do our best to reconstruct waveforms. If not, the reason why we have to use auditory models becomes ambiguous.

Thus, I will point out some candidates of objective criteria concerning

0) how we construct a faithful auditory model,
1) what physical values are modified in each stage of the model,
2) what physical values are emphasized through the whole model, and
3) whether the physical values through the model are useful for a certain application.

It is significant to evaluate the output of the model before rashly applying it to any applications.

3. What method or model can treat an entity of 200 ms or more?

It seems common that physical values analyzed from speech waves are different from the psychophysical results perceived from those speech waves. Therefore, we have to construct a model dealing with not physical values but psychophysical values. Dr. Ghitza introduced an interesting paradigm. He showed us the difference between physical and psychological distances and an implementation of psychophysical measurements into a learning method for the parameters of a speech recognizer using diphons. This is one of the useful trials for filling up gaps between physical and psychophysical values.

As an example, how do we deal with the relations between objective phonemes and pre- and/or post-phonemes? Previous research[1][2] has shown that psychophysical values are very different from physical values, as shown in the Fig. 1 case in which the physical values of the shadowed areas are the same, but the psychophysical results shift toward the opposite side of the adjacent phonemes by contextual effects. However, I have not seen recent analysis methods to treat these phenomena.

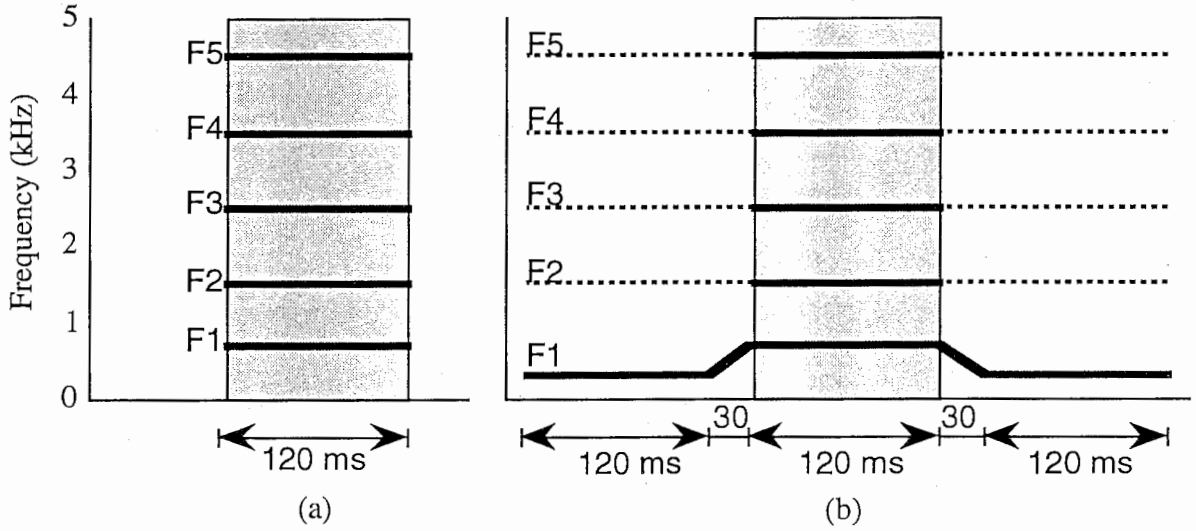


Figure 1. (a) Isolated synthesized vowel and (b) synthesized vowel with pre- and post-phonemes (vowel: filled line and broken line, and single-formant: filled line only). Subjects perceived the F1 of (b) higher than that of (a), when the F1 of the adjacent phonemes are lower than that of the central vowel.

Until now, analysis of speech waves for speech recognition and coding has handled each short term frame, in which the characteristics of the speech wave are assumed to be stable. The length of the frame is 20 ~ 30 ms. Thus, if we model the above article in this situation, we have to represent relations of inter-frames. HMM might be a typical method and describe such relations implicitly. But, we can not observe how HMM modifies differences between physical and psychological values.

Thus, we have to look for other methods to describe the relations of intra-frames covering two or more phoneme lengths (200 ms or more) explicitly, and we have to construct new useful models for applications such as speech recognition and coding. The psychophysical effects which appear in such a frame length are masking, the contextual effect between phonemes, etc. Some researchers are studying these articles[3][4].

4. Relations between temporal fine structures and sound segregation

Let us consider the modeling of sound segregation. Previous research has indicated that sound onset/offset is a significant cue for sound segregation. Dr. Carlyon and Prof. Hartmann showed us that difference in fundamental frequencies (F0) and the relations of their harmonics are also significant cues. Dr. Cooke introduced a speech enhancement model in a noisy environment using onset/offset, and the relations of F0 harmonics extracted in a sound-spectrogram. Other models for the same purpose are also almost totally formulated in sound spectrograms.

This raises a question. Why did we use only sound spectrograms? Sound spectrograms are usually calculated by using the DFT, or a squared sum of filter bank output values and the DFT, or the squared sum of filter bank output values is one of the averaging operations in each analyzed frame. Thus, as Dr. Patterson pointed out, it happens that the perception of damped and ramped sinusoids is different even though their long term sound spectrograms are the same. This finding suggests that temporal fine structures are significant and that the sound spectrogram reduces such information. Additionally, Prof. Yost showed that the perception of iterated ripple noises is the same when the first peaks of the autocorrelation function are the same, although their sound spectrograms are different. This finding also suggests that the sound spectrogram is not useful for constructing auditory models.

On the other hand, there is much useful data in the time domain. For example, let us assume $s(t)$ as the output of a band-pass filter (BPF) of a filterbank when two sounds go through it simultaneously,

$$\begin{aligned} s(t) &= A(t)\sin\omega t + B(t)\sin(\omega t + \phi) \\ &= C(t)\sin(\omega t + \theta(t)) \\ C(t) &= \sqrt{A^2(t) + 2A(t)B(t)\cos\phi + B^2(t)} \\ \theta(t) &= \arctan\left(\frac{B(t)\sin\phi}{A(t) + B(t)\cos\phi}\right), \end{aligned}$$

where $A(t)$ and $B(t)$ are modulated amplitudes of two sounds, ω is the center frequency of the BPF, and ϕ is the phase difference between the two sounds. In the equation, $C(t)$ is related to sound spectrograms and we almost use $C(t)$ only for sound segregation. However, $\theta(t)$ also has much information about each sound. For example, the peak sequences of $s(t)$, t_n and t_{n+1}, are deviated as a function of $A(t)$ and $B(t)$, and the distance between the two peak positions is

$$\begin{aligned} \tan\left[\omega(t_{n+1} - t_n) - 2\pi\right] &= \tan\left[\theta(t_{n+1}) - \theta(t_n)\right] \\ &= \frac{(A_{n+1}B_n - A_nB_{n+1})\sin\phi}{A_nA_{n+1} + (A_{n+1}B_n + A_nB_{n+1})\cos\phi + B_nB_{n+1}} \end{aligned}$$

where $A_n = A(t_n), A_{n+1} = A(t_{n+1}), B_n = B(t_n)$ and $B_{n+1} = B(t_{n+1})$.

These might be related to the deviation from "phase-lock" and we can reconstruct $A(t)$ and $B(t)$ from $s(t)$ by using these equations[5][6].

Thus, as Prof. Patterson pointed out, we should also pay attention to temporal structures for the modeling of not only sound segregation but other interesting psychophysical and physiological things.

5. Conclusion

Until now, I have never heard that any psychophysical and physiological models of human audition work better than engineering techniques such as DFT, LPC

and Cepstrums, except for some models of auditory periphery in noisy environments. However, recentl development of engineering techniques seems to be stagnant. Investigations about HMM also seem to have seen their golden days.

Thus, we have to find new breakthroughs in speech processing technology. I believe that knowledge and findings from psychophysics and physiology will help us to overcome this recent stagnation.

References

[1] Akagi, M : "Modeling of contextual effects based on spectral peak interaction", J. Acoust. Soc. Am., 93, 2, 1076-1086 (1993)

[2] Akagi, M., van Wieringen, A., and Pols, L. C. W.: "Perception of central vowel with pre- and post-anchors", Proc. ICSLP-94, S10-7 (1994)

[3] Akagi, M. : "Psychoacoustic evidence for contextual effect models", Speech Perception, Production and Linguistic Structure (Tohkura, Y., Vatikiotis-Bateson, E., and Sagisaka, Y. Eds.), Ohmsha Tokyo and IOS Press Amsterdam, 62-78 (1992)

[4] Aikawa, K. and Saito, T. : "Noise robust speech recognition using a dynamic-cepstrum", Proc. ICSLP-94, S26-26 (1994)

[5] Igawa, H. : "A fundamental study on modeling of auditory segregation based on phase deviation", Japan Advanced Institute of Science and Technology, Master Thesis (1994).

[6] Igawa, H. and Akagi, M. : "Modeling of Auditory Segregation", ASJ Fall meeting, 3-4-15 (1994).

What is needed in the computational approach to auditory perception?

Makio Kashino

NTT Basic Research Laboratories
3-1, Morinosato Wakamiya, Atsugi
Kanagawa 243-01, JAPAN
E-mail: kashino@av-hp.ntt.jp

The computational approach, which has been a fairly successful contributor to the field of vision research during the past 10 years or so, is being introduced into the field of hearing research. In this ATR workshop, several pioneering studies were reported along this line. I would like to make some comments on what I believe is needed in the computational approach to auditory perception.

1. What is the computational approach to perception?

The word "computational" has been assigned various meanings in hearing research. Some people have used the term to refer to "a model described by mathematical formulas," while others use it to refer to "a simulation using a computer." However, David Marr, who established the computational approach in vision, used the term in a more restricted sense (Marr, 1982). He distinguished three levels in the understanding of information processing. "Computational theory" is the first level, which clarifies what is computed and why in a given process. In other words, the computational theory specifies inputs (givens), outputs (goals), and constraints to yield a unique mapping function between the input and output. The second level is a level of "representation and algorithm," which chooses a representation for the input and output and a procedure for transforming one into the other. The final level involves the "hardware implementation" of the representation and algorithm.

According to this distinction, most "computational" studies conducted in hearing research are not actually based on computational theory. Rather, they are at the level of algorithm or implementation. In other words, they are more "experiments of performance" than "theories of competence" of the auditory system. I do not mean to suggest that studies of algorithm or implementation are without merit. Indeed, they are important because they enable us to examine otherwise unrelated data using constructive methods. I would simply like to point out that experiments of performance are not sufficient to understand information processing. For example, suppose we were trying to understand the operation of an AM radio receiver. We find a variable capacitor and a coil in it, and we measure them. By analyzing the connections, we also find that these two devices make a variable bandpass filter. Finally, we successfully program on a computer a "front-end model of a radio receiver" that works in exactly the same way the actual one does. However, this is not enough. To understand what the variable filter is for,

we have to know how AM radio broadcasting works. Then, we would be able to understand that it is not essential to use a variable capacitor or a variable induction coil, and we would be able to predict that a rectifier should follow the filter to extract the amplitude envelope.

To return to the main topic: Most current auditory models are not based on computational theories of auditory perception, but on physiological or psychophysical data. To simulate performance is one thing; to understand competence is quite another.

2. The goal of auditory perception

The first step toward understanding the competence of the auditory system is to clarify its goal. Marr (1982) stated that the final goal of the visual system is to recover the three-dimensional structure of the real world from the two-dimensional retinal image. Many systematic studies have been generated by assuming that each sub-process of the visual system should be appropriate for this purpose. Now, what is the goal of the auditory system?

The answer may be to determine from the acoustic signal reaching our ears, what acoustic events are occurring and where. Recent researchers on "auditory scene analysis" seem to recognize this point well (Bregman, 1990). However, it is not well understood what kinds of sub-processes are required to achieve this final goal. Studies of auditory scene analysis have focused on the problem of how to group and segregate frequency components when two or more acoustic events are taking place simultaneously as is often the case in the real world. This line of research may be able to address the problem of restoring missing portions of acoustic signals. However, other problems, such as perceptual constancy for example, have received little attention. The problem is how to recover sound source information despite the fact that the acoustic signal reaching the ears is modified considerably by room acoustics, telephone frequency response, and so on. Additionally, the identification of acoustic events (recovering the method of producing the sound, the material, size, shape, etc.) is also one of the important but unexamined roles of the auditory system. The first step of the computational approach to auditory perception is to formalize such problems.

3. Constraints of auditory perception

Auditory perception may be thought of as a process of solving inverse problems of acoustics. Most such inverse problems are ill-posed problems in which the solution is not unique. Despite this, we usually experience only a single percept, suggesting that our auditory system does its computation using some implicit assumptions (constraints). To regularize ill-posed problems by imposing constraints is equivalent to minimizing some cost functions (Poggio et al., 1985). The question then becomes: what kinds of constraints are used by the auditory system?

One may think that the auditory system uses laws that hold for acoustic events in general (ecologically validity) as constraints. If so, we need to know in detail about the physical nature of acoustic events in the real world

(Richards, 1988). In studies of auditory scene analysis, several heuristics of grouping frequency components have been proposed. Though some of them intuitively appear ecologically valid, they are not based on a rigorous analysis of acoustic events, nor are they based on a theoretical analysis of their potential to regularize ill-posed problems.

4. Problems of representation

In the computational approach, the representation of information is essential. For instance, though both Roman and Arabic numerals can represent numbers, Arabic numerals are much more convenient in the process of multiplication. I would now like to comment on problems of representation in current auditory research.

First, it is rare for the problems of representation to be treated theoretically. For example, there are many studies of frequency "representation" in the auditory periphery, but most of them are a simulation of physiological or psychophysical data. Researchers seldom ask why information must be represented in a specific way. It is important to examine what aspects of the acoustic signal include important features of acoustic events, or what kinds of information should be explicitly represented to make the following processes easy. Specifically, the process of representing dynamic aspects of acoustic events will be a central topic of auditory perception. In this workshop, it was intriguing to see several speakers report psychophysical evidence indicating that the auditory system actually uses temporal representations in the perception of pitch, timbre, and auditory entity formation.

Second, relatively naive "isomorphism" is sometimes found in the discussion of representation. For example, one may assume that in order to perceive a continuous tone there must be isomorphic (in this case, continuous) neural activity somewhere in the auditory system. However, the same information can be represented equally by transitional activities of onset and offset. Representation of an acoustic event does not have to be isomorphic to the acoustic event. Indeed, isomorphic representations would be poor from the viewpoint of information processing, because, generally speaking, the input acoustic signal is imperfect, and at the same time, redundant. In order to obtain stable representations of acoustic events, it is necessary to restore missing portions taking advantage of signal redundancy, or to reduce redundancy to represent necessary information effectively. Therefore, when looking for auditory representation, we should analyze what kind of information is necessary to the auditory system (to be represented explicitly), and how the necessary information is coded effectively.

Acoustic signals reaching the ears reflect not only characteristics of the source events, but also various factors such as transducer characteristics (room, telephone, etc.), sound source position, and movement by the listener. If the final goal of the auditory system is to obtain an invariant description of the acoustic events, various contaminating factors included in the input signal should be factored out during the course of processing. In vision, Marr (1982) assumed that a similar task is achieved through two stages of processing. First, appropriate representations of image structures

and changes (primal sketches) are obtained. Second, many parallel processes operate on the primal sketches to get viewer-centered representations of the geometrical structures of visible surfaces (2 1/2-D sketch). Finally, 3-D representations of objects in an object-centered coordinate frame are obtained. It would be important to find the framework applicable to auditory perception. Again, we should be familiar with the real-world acoustics.

5. Computational theory and empirical sciences

Computational analysis is useful in locating psychophysical and physiological studies in the attempt to understand the total system. At the same time, we should note that the computational approach may not lead to a perfect understanding of the actual biological systems, because biological systems, which have been developed through evolution, may not always be mathematically optimal. It may be more appropriate to think of the biological perceptual systems as a "bag of tricks" that can compute roughly appropriate answers quickly (Ramachandran, 1990). Therefore, empirical tests are crucial to our understanding of the biological perceptual systems. However, biological systems cannot be completely unlawful. Therefore, computational analysis is useful in understanding why perception is possible at all.

6. Conclusion

In order to understand information processing in the auditory system, it is necessary to formalize the problems to be solved by the system, and to analyze why those problems can be solved. In other words, we need theories of competence for the auditory system.

References

Marr, D. (1982). Vision. New York: Freeman.
Poggio, T., Torre, V. & Koch, C. (1985). Computational vision and regularization theory. Nature, 317, 314-319.
Ramachandran, V. S. (1990). Visual perception in people and machines. In A. Blake & T. Troscianko (Eds.), AI and the Eye. New York: Wiley.
Richards, W. (Ed.). (1988). Natural Computation. Cambridge: MIT Press.

List of Participants

ABE Yuzo ABE@YDI-01.YD.HM.RD.SANYO.CO.JP
SANYO Electric Co., Ltd., Hyper Media Res. Center Tel: +81-6-900-3517
Image and Audio Dept. Fax: +81-6-900-3557

AIKAWA Kiyoaki aik@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1029
Fax: +81-7749-5-1008

AKAGI Masato akagi@jaist.ac.jp
School of Information Science Tel: +81-761-51-1236
Japan Advanced Institute of Science and Technology, Hokuriku Fax: +81-761-51-1341

AMOUYAL Jerome jerome@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1012
Fax: +81-7749-5-1008

AOKI Shigeaki aoki@nttspch.ntt.jp
NTT Human Interface Labs. Tel: +81-468-59-3653
Fax: +81-468-55-1054

BIEM Alain biem@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1082
Fax: +81-7749-5-1008

BROWN Guy g.brown@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1026
(Dept. of Computer Science, Univ. of Sheffield) Fax: +81-7749-5-1008

CARLYON Robert P. bob.carlyon@mrc-apu.cam.ac.uk
MRC Applied Psychology Unit Tel: +44-223-355294 ext. 720
Fax: +44-223-359062

CHISAKI Yoshifumi chisaki@eecs.kumamoto-u.ac.jp
Kumamoto Univ. Tel: +81-96-344-2111
Fax: +81-96-345-1553

COLLOMB Alexis xcollomb@itl.atr.co.jp
ATR Interpreting Telecom. Res. Lab. Tel: +81-7749-5-1334
Fax: +81-7749-5-1308

COOK Norman D. cook@res.kutc.kansai-u.ac.jp
Faculty of Informatics Tel: +81-726-90-2447
Kansai Univ. Fax: +81-726-90-2493

COOKE Martin M.Cooke@dcs.shef.ac.uk
Dept. of Computer Science Tel: +44-742-768555
Univ. of Sheffield Fax: +44-742-780972

CRAWFORD Malcolm mmalc@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1089
(Dept. of Computer Science, Univ. of Sheffield) Fax: +81-7749-5-1008

DANG Jianwu dan@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1028
Fax: +81-7749-5-1008

DERMODY Phillip NAL@vaxa.ee.su.OZ.AU
Speech Communication Section Tel:
National Acoustic Labs. Fax: +61-2-411-8273

EBATA Masanao ebata@eecs.kumamoto-u.ac.jp
Dept. of Electrical Engineering and Computer Science Tel: +81-96-344-2111
Kumamoto Univ. Fax: +81-96-345-1553

EIGSTI Inge-Marie eigsti@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1088
Fax: +81-7749-5-1008

ERICKSON Donna erickson@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1042
Fax: +81-7749-5-1008

FAYCAL Zitouni zitouni@atr-sw.atr.co.jp
ATR Communication Sys. Res. Labs. Tel: +81-7749-5-1291
Fax: +81-7749-5-1208

FERNALD Anne sjcgw!fernald@kuis.kyoto-u.ac.jp
Stanford Japan Center-Research Tel: +81-75-752-7073
Fax: +81-75-752-1120

FUJIMURA Osamu
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1050
(Dept. of Speech & Hearing Science, Ohio State Univ.) Fax: +81-7749-5-1008

FUJITA Satoru fujita@hip.atr.co.jp
ATR Human Information Processing Res. Lab. Tel: +81-7749-5-1054
Fax: +81-7749-5-1008

FUKUNISHI Kohyu fukunisi@harl.hitachi.co.jp
Advanced Res. Lab. Tel: +81-492-96-6111
Hitachi Ltd. Fax: +81-492-96-6006

FUNASAKA Sotaro No
Dept. of Otolaryngology Tel: +81-3-3342-6111
Tokyo Medical College Fax: +81-3-3346-9275

GHITZA Oded og@research.att.com
Acoustics Res. Dept. Tel:
AT&T Bell Labs. Fax: +1-908-582-7308

GRACCO Vincent L. GRACCO%LENNY@YALEVMS.BITNET
Haskins Labs. Tel:
Fax:

GREEN Phil P.Green@dcs.shef.ac.uk
Dept. of Computer Science Tel: +44-742-768555
Univ. of Sheffield Fax: +44-742-780972

HABARA Kohei
ATR International Tel: +81-7749-5-1112
Fax: +81-7749-5-1109

List of Participants

HARTMANN William Morris
Dept. of Physics
Michigan State Univ.
MARTMANN@msupa.pa.msu.edu
Tel: +1-517-355-5202
Fax:

HASEGAWA-JOHNSON Mark
M.I.T. Res. Lab. of Electronics
johnson@lexic.mit.edu
Tel: +1-617-253-5957
Fax: +1-617-277-0126

HERMANSKY Hynek
Oregon Graduate Institute
hynek@eeap.ogi.edu
Tel: +1-503-690-1136
Fax: +1-503-690-1334

HIRAHARA Tatsuya
Hearing Science Res. Group
NTT Basic Res. Labs.
hirahara@siva.ntt.jp
Tel: +81-462-40-3610
Fax: +81-462-40-4725

HIRAI Hiroyuki
ATR Human Information Processing Res. Lab.
hirai@hip.atr.co.jp
Tel: +81-7749-5-1053
Fax: +81-7749-5-1008

HOHMANN Voker
FB 8 / Physics
Univ. of Oldenburg
vh@hinz.physik.uni-oldenburg.de
Tel: +49-441-798-5468
Fax: +49-441-798-3698

HONDA Masaaki
Information Science Res. Lab.
NTT Basic Res. Labs.
hon@av-sun2.NTT.jp
Tel: +81-462-40-3580
Fax: +81-462-40-4721

HONDA Kiyoshi
ATR Human Information Processing Res. Lab.
honda@hip.atr.co.jp
Tel: +81-7749-5-1051
Fax: +81-7749-5-1008

HORIKAWA Junsei
Medical Res. Institute
Tokyo Medical and Dental Univ.
F00733@sinet.ad.jp
Tel: +81-3-5280-8050
Fax: +81-3-5280-8073

HOSOI Yuji
The Dept. of Otorhinolaryngology
School of Medicine, Kinki Univ.
No
Tel: +81-723-66-0221
Fax: +81-723-66-0206

IRINO Toshio
Information Science Res. Lab.
NTT Basic Res. Labs.
irino@av-sun2.ntt.jp
Tel: +81-462-40-3597
Fax: +81-462-40-4721

ITAKURA Fumitada
Nagoya Univ.
ita@itakura.nuee.nagoya-u.ac.jp
Tel: +81-52-789-3171
Fax: +81-52-789-3172

IWASAWA Hideki
Dept. of Psychology
College of Humanities & Sciences, Nihon Univ.
iwasawa@chsgw.chs.nihon-u.ac.jp
Tel: +81-3-3329-1151
Fax: No

JAMIESON Donald G.
Hearing Health Care Res. Unit
Univ. of Western Ontario
Jamieson@uwovax.uwo.ca
Tel: +1-519-661-3901
Fax: +1-519-661-3805

List of Participants

KABURAGI Tokihiko
Information Science Res. Lab.
NTT Basic Res. Labs.
kabu@voice-sun.NTT.jp
Tel: +81-462-40-3643
Fax: +81-462-40-4721

KAJITA Shoji
School of Engineering
Nagoya Univ.
kaji@itakura.nuee.nagoya-u.ac.jp
Tel: +81-52-789-3626 (or 4432)
Fax: +81-52-789-3172

KAMADA Tsutomu
Dept. of Oral Physiology
Hokkaido Univ., School of Dentistry
kamada@den.hines.hokudai.ac.jp
Tel: +81-11-706-4230
Fax: +81-11-706-4919

KASHINO Makio
Hearing Science Res. Group
Information Science Res. Lab., NTT Basic Res. Labs.
kashino@ar-hp.ntt.jp
Tel: +81-462-40-3627
Fax: +81-462-40-4725

KASHINO Kunio
Dept. of Electrical Engineering
Faculty of Engineering, Univ. of Tokyo
kashino@mtl.t.u-tokyo.ac.jp
Tel: +81-3-3812-2111
Fax: +81-3-5800-6922

KASUYA Hideki
Dept. of Electrical and Electronic Engr.
Utsunomiya Univ.
kasuya@utsunomiya-u.ac.jp
Tel: +81-286-61-3401 ext. 440
Fax: +81-286-89-0971

KATO Hiroaki
ATR Human Information Processing Res. Lab.
kato@hip.atr.co.jp
Tel: +81-7749-5-1022
Fax: +81-7749-5-1008

KAWAHARA Hideki
ATR Human Information Processing Res. Lab.
kawahara@hip.atr.co.jp
Tel: +81-7749-5-1020
Fax: +81-7749-5-1008

KAWATO Mitsuo
ATR Human Information Processing Res. Lab.
kawato@hip.atr.co.jp
Tel: +81-7749-5-1040
Fax: +81-7749-5-1008

KITAZAWA Shigeyoshi
Dept. Computer Science
Faculty of Engineering, Shizuoka Univ.
Kitazawa@cs.shizuoka.ac.jp
Tel: +81-53-471-1171
Fax: +81-53-475-4595

KOMAKINE Takashi
Akita Res. Institute of Advanced Technology
KGE02166@niftyserve.or.jp
Tel: +81-188-66-5800
Fax: +81-188-66-5803

KURAKATA Kenji
Faculty of Human Sciences
Osaka Univ.
MXC01642@niftyserve.or.jp
Tel: +81-6-850-5666
Fax: +81-6-850-5667

KUSAKAWA Naoki
ATR International
Tel: +81-7749-5-1195
Fax: +81-7749-5-1008

LENZO Kevin A.
ATR Interpreting Telecom. Res. Lab.
lenzo@itl.atr.co.jp
Tel: +81-7749-5-1344
Fax: +81-7749-5-1308

List of Participants

MAGNUSON James
ATR Human Information Processing Res. Lab.
magnuson@hip.atr.co.jp
Tel: +81-7749-5-1084
Fax: +81-7749-5-1008

MASAKI Shinobu
ATR Human Information Processing Res. Lab.
masaki@hip.atr.co.jp
Tel: +81-7749-5-1001
Fax: +81-7749-5-1008

MASUDA Ikuyo
Central Res. Labs.
Matsushita Electric Industrial Co., Ltd.
ikuyo@crl.mei.co.jp
Tel: +81-7749-8-2530
Fax: +81-7749-8-2578

MATSUI Michinao
Graduate School of Language and culture
Osaka Univ.
matui@lisa.lang.osaka-u.ac.jp
Tel: +81-6-850-6111
Fax: No

MATSUOKA Takahide
Dept. of Electrical and Electronic Engr.
Utsunomiya Univ.
matsuoka@al.cc.utsunomiya-u.ac.jp
Tel: +81-286-61-3401 ext. 415
Fax: +81-286-63-2726

McADAMS Stephen
Laboratoire de Psychologie Experimentale(CNRS)
Universite Rene Descartes
smc@nadia.ircam.fr
Tel: +33-1-40519842
Fax:

MEDDIS Ray
Dept. of Human Sciences
Loughborough Univ.
R.Meddis@lut.ac.uk
Tel: +44-509-230452
Fax:

MEKATA Tsuyoshi
Central Res. Labs.
Matsushita Electric Industrial Co., Ltd.
mekata@crl.mei.co.jp
Tel: +81-7749-8-2530
Fax: +81-7749-8-2578

MUNHALL Kevin
ATR Human Information Processing Res. Lab.
munhall@hip.atr.co.jp
Tel: +81-7749-5-1016
Fax: +81-7749-5-1008

NAKAMURA Satoshi
Nara Institute of Science and Technology
Tel: +81-7437-2-5111
Fax:

NEY Hermann
Aachen Univ. of Technology
ney@informatik.rwth-aachen.de
Tel: +49-421-8021600
Fax: +49-421-8888218

OBARA Kazuaki
Central Res. Labs.
Matsushita Electric Industrial Co., Ltd.
obara@crl.mei.co.jp
Tel: +81-7749-8-2522
Fax: +81-7749-8-2577

ODA Masaomi
ATR Human Information Processing Res. Lab.
oda@hip.atr.co.jp
Tel: +81-7749-5-1033
Fax: +81-7749-5-1008

OHGUSHI Kengo
Kyoto City Univ. of Arts
Faculty of Music
JAH03424@niftyserve.or.jp
Tel: +81-75-332-0701
Fax: +81-75-332-0709

List of Participants

OHNISHI Michihiro
ATR Human Information Processing Res. Lab.
moohishi@hip.atr.co.jp
Tel: +81-7749-5-1038
Fax: +81-7749-5-1008

OHYAMA Ghen
Dept. of Otolaryngology
School of Medicine, Univ. of Tokushima
ohyama@clin.med.tokushima-u.ac.jp
Tel: +81-886-33-7168 (or 7169)
Fax: +81-886-33-7170

OKUNO Hiroshi
Information Science Res. Lab.
NTT Basic Res. Labs.
okuno@nuesun.ntt.jp
Tel: +81-462-40-3646
Fax: +81-462-40-4708

PATTERSON Roy
ATR Human Information Processing Res. Lab.
(MRC Applied Psychology Unit)
roy.patterson@mrc-apu.cam.ac.uk
Tel: +81-7749-5-1002
Fax: +81-7749-5-1008

RIQUIMAROUX Hiroshi
Neural Systems Lab., Frontier Res. Program
The Institute of Physical and Chemical Res., RIKEN
rikimaru@murasaki.riken.go.jp
Tel: +81-48-462-1111
Fax: +81-48-462-4698

SASAKI Takayuki
Miyagi Gakuin Women's College
g26484@cctu.cc.tohoku.ac.jp
Tel: +81-22-277-8374
Fax: +81-22-277-6186

SATO Masaaki
ATR Human Information Processing Res. Lab.
masaaki@hip.atr.co.jp
Tel: +81-7749-5-1039
Fax: +81-7749-5-1008

SCHROETER Juergen
Information Principles Res.
Acoustics Res. Dept., AT&T Bell Lab.
jsh@research.att.com
Tel: +1-908-582-7059
Fax: +1-908-582-7308

SHIKANO Kiyohiro
Nara Institute of Science and Technology
shikano@is.aist-nara.ac.jp
Tel: +81-7437-2-5280
Fax: +81-7437-2-5289

SINGER Harald
ATR Interpreting Telecom. Res. Lab.
singer@itl.atr.co.jp
Tel: +81-7749-5-1389
Fax: +81-7749-5-1308

SLANEY Malcolm
Interval Res. Inc.
malcolm@interval.com
Tel:
Fax:

SOQUET Alain
Insitut des Langues Vivantes et de Phonetique
asoquet@ulb.ac.be
Tel: +32-2-650-36-60
Fax: +32-2-650-20-07

STERN Richard M.
Dept. of Electrical and Computer Engineering
Carnegie Mellon Univ.
rms@SPEECH1.CS.CMU.EDU
Tel: +1-412-268-2535
Fax: +1-412-268-3890

TANAKA Masako
ATR Human Information Processing Res. Lab.
masako@hip.atr.co.jp
Tel: +81-7749-5-1025
Fax: +81-7749-5-1008

List of Participants

TANIGUCHI Ikuo
G01130@sinet.ad.jp
Dept. of Neurophysiology, Medical Res. Institute
Tokyo Medical and Dental Univ.
Tel: +81-3-5280-8073
Fax: +81-3-5280-8073

TATENO Takashi
NTT Basic Res. Lab.
Tel:
Fax: +81-462-40-4725

TIEDE Mark
tiede@hip.atr.co.jp
ATR Human Information Processing Res. Lab.
Tel: +81-7749-5-1083
Fax: +81-7749-5-1008

TOHKURA Yoh'ichi
tohkura@hip.atr.co.jp
ATR Human Information Processing Res. Lab.
Tel: +81-7749-5-1000
Fax: +81-7749-5-1008

TSUKAMOTO Taeko
PFD03327@niftyserve.or.jp
Dept. of Neurology and Psychiatry
Kobe Univ. School of Medicine
Tel: +81-742-47-6472
Fax: +81-742-47-6472

TSUZAKI Minoru
tsuzaki@hip.atr.co.jp
ATR Human Information Processing Res. Lab.
Tel: +81-7749-5-1021
Fax: +81-7749-5-1008

UEDA Kazuo
h50015@sakura.kudpc.kyoto-u.ac.jp
Faculty of Letters
Kyoto Prefectural Univ.
Tel: +81-75-781-3131
Fax: +81-75-781-1841

UNO Yoji
uno@hip.atr.co.jp
ATR Human Information Processing Res. Lab.
Tel: +81-7749-5-1041
Fax: +81-7749-5-1008

USAGAWA Tsuyoshi
tuie@eecs.kumamoto-u.ac.jp
Kumamoto Univ.
Tel: +81-96-344-2111 ext. 3622
Fax: +81-96-345-1553

V-BATESON Eric
bateson@hip.atr.co.jp
ATR Human Information Processing Res. Lab.
Tel: +81-7749-5-1057
Fax: +81-7749-5-1008

YAMADA Masashi
PFE03077@niftyserve.or.jp
Dept. of Musicology
Osaka Univ. of Arts
Tel: +81-721-93-3781
Fax: +81-721-93-5587

YAMADA Yoshinori
yamada@crl.mei.co.jp
Central Res. Labs.
Matsushita Electric Industrial Co., Ltd.
Tel: +81-7749-8-2530
Fax: +81-7749-8-2578

YAMADA Reiko
yamada@hip.atr.co.jp
ATR Human Information Processing Res. Lab.
Tel: +81-7749-5-1024
Fax: +81-7749-5-1008

YAMADA Takeshi
Nara Institute of Science and Technology
Tel: +81-7437-2-5111
Fax:

List of Participants

YEHIA Hani Camille hani@itakura.nuee.nagoya-u.ac.jp
Itakura Lab., Dept. of Electronic Information Tel: +81-52-789-4432
School of Engineering, Nagoya Univ. Fax: +81-52-789-3172

YOST William A. wyost@wpo.it.luc.edu
Parmly Hearing Institute Tel: +1-312-508-2710
Loyola Univ. of Chicago Fax: +1-312-508-2719