TR－H－016

Time-domain comb filtering for speech separation

A. de Cheveigné

1993. 7. 27

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　☎07749-5-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

Time-domain comb filtering for speech separation

A. de Cheveigné

1. Introduction

The auditory system uses differences that occur in the harmonic structure of concurrent sounds, such as speech, to separate them. This is one aspect of what is known as the "cocktail-party effect".

Several models have been proposed to explain how this is done (see de Cheveigné 1993 for a review). They usually assume that signals to be separated are purely harmonic. Psychoacoustic and physiological experiments designed to test them likewise employ such stimuli. However real speech is often very imperfectly harmonic, and it is not clear how well the models will work in that case.

In order to determine how well a model can perform its task on "real" speech, I implemented its basic processing scheme as a front-end to a speech recognition system and measured the effect on the rates in a recognition task. To the extent that this processing reflects that of the perception model, and that the task is typical of the perception of speech in "real" situations, the results should give some indication of the plausibility of the model.

It is stressed that the aim is <u>not</u> to develop a speech separation system. The results might however be of some use in designing such a system. I also do not wish to reproduce quantitatively the recognition rates obtained in psychoacoustic experiments. To do so would require postulating many details of the physiological implementation, and thus obscuring the essential features of the model. Instead, I wish to find out if its processing principle, implemented in some form, can be effective in tasks typical of the "real world".

1.1. A physiological model of time-domain cancellation

Various schemes for harmonic sound separation have been proposed, both as perception models and signal processing methods. Most of them assume some form of frequency analysis, but it is also possible to imagine harmonic sound separation in the time domain, for example using the neural equivalent of a time-domain comb-filter (de Cheveigné, 1993). Fig. 1 shows such a filter applied to one channel (group of fibers with similar characteristics) of the auditory-nerve. The characteristics of the gating neuron are such that it lets pass every spike that arrives along the direct path, unless a spike arrives simultaneously along the delayed path. Simulation of this filter with data recorded in the guinea-pig auditory-nerve shows that such processing can be effective in separating the correlates of each individual vowel from the neural response to a mixture of the two.

An array of such filters, all tuned to the same lag, might process the entire auditory-nerve response pattern. The effect of the neural filter is similar (though not equivalent) to that of a time-domain comb-filter operating on a linear representation. If we model it in this way, and neglect other sources of non-linearity, we can invert the order of the filtering and replace the array of filters with a single filter preceding cochlear filtering and operating on a linear representation of the signal (Fig. 2). We can consider that this simple comb-filter represents (at a certain level of abstraction) the time-domain mechanism of our physiological model.

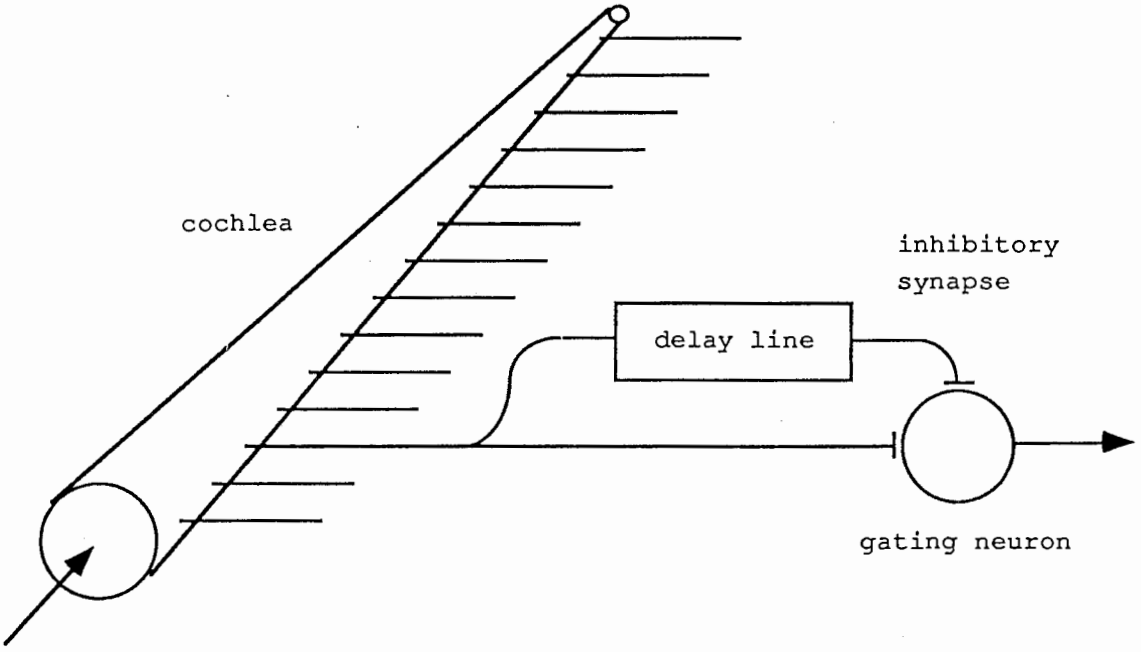

Fig. 1. Neural filter for harmonic sound cancellation. The delay is adjusted to equal the period of the interference. Any spike preceded by another spike at that delay is eliminated from the spike train.

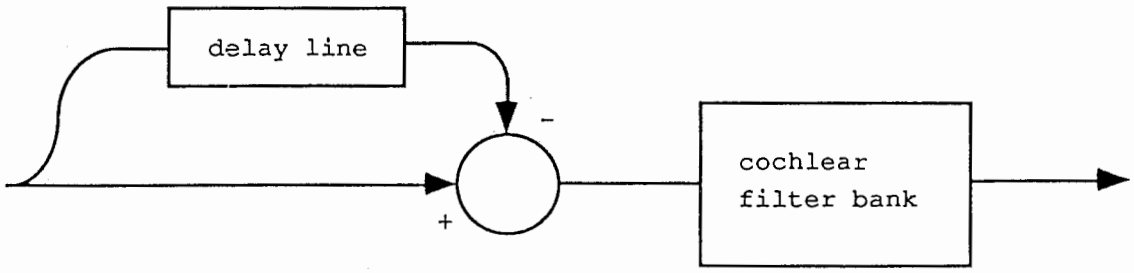

Fig. 2. Equivalent model operating on a linear representation of sound.

The effect of a cancellation-type comb filter is easy to understand in the time-domain. If the period of an interfering signal equals the lag (delay) parameter of the filter, it is in effect subtracted from itself and the output is zero. It can be also understood in the frequency domain by remarking that the comb filter has a series of zeros equally spaced at multiples of the inverse of the lag parameter (Fig. 3). If the components of the interference coincide with that series, then it is canceled.

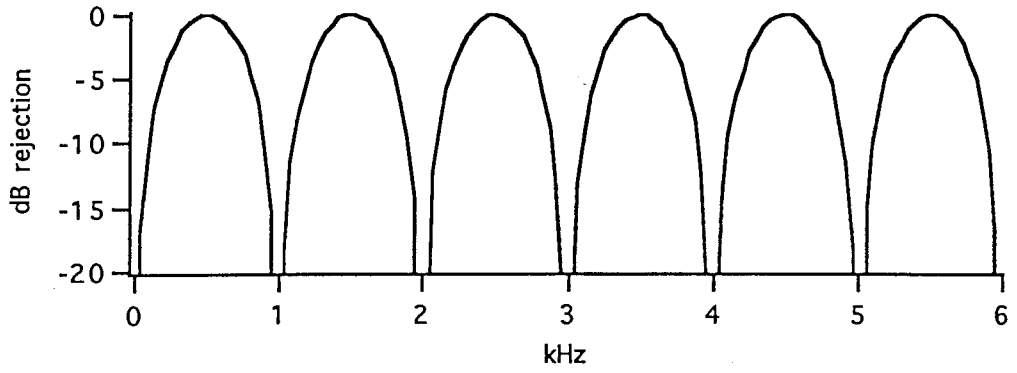

Fig. 3. Transfer function of a comb filter of impulse response $h(t) = (\delta(t) - \delta(t-L))/2$, with L = 1 ms.

1.2. Questions

As other models, this model assumes that speech is perfectly periodic, and it is thus not certain how well it can cope with the aperiodicity of real speech. If the interference is not perfectly periodic, it won't be perfectly canceled. In addition, filtering may cancel useful information within the target signal.

In this work, I examine the following questions:

1) Is time-domain cancellation effective applied to real speech, and to what extent ?

2) What are the relative impacts of the two factors that limit effectiveness: interference cancelation residue and spectral distortion of the target ?

3) Is it possible reduce the effects of spectral distortion by applying similar distortion to reference templates?

4) Is time-domain processing better, in some sense, than frequency-domain processing?

5) Is harmonic cancellation of interference better, in some sense, than harmonic enhancement of the target?

I also investigate several other issues, such as resolution of fundamental frequency resolution, and the choice of feature format (linear spectrum vs log spectrum or cepstrum).

2. Methods

2.1. Principle

The effects of interference and interference reduction processing were assessed by measuring their effects on the recognition scores of a speech recognition system.

2.2. Recognition task and method

The task was to recognize words belonging to a set of 100, by comparison with reference templates. The speech recognizer used DTW pattern-matching between arrays of feature vectors, using standard Euclidian distance. Features were based on 128 coefficient linear magnitude spectra calculated using a 256-point Hanning-window at a 128 sample frame-rate. Each 128 coefficient spectrum was condensed to a 16 coefficient vector by averaging samples 8 by 8.

The results presented here are essentially based on this representation, but others were also tried, in particular a 16-coefficient FFT cepstrum with "Tohkura weighting" (Tohkura 1987).

2.3. Target, reference, and interference database

The database consisted of one hundred short Japanese words taken from the ATR database (Kuwabara et al. 1989). Speech data were sampled at 12 kHz, 16 bits resolution. "Silent" portions were eliminated, based on labels. The same word set was used for targets and for reference templates, and also as interference.

Each target word was paired with another word from the word set chosen at random. Each word of the set served exactly once as target and once as interference. Words were of similar duration and started together, and the degree of overlap of their signals was therefore high. The components were added at signal-to-noise ratios (SNR) of 6, 0, -6 and 12 dB. SNR was defined globally and implemented by applying a fixed factor to either component in a pair. No attempt was made to control the SNR within each individual pair. Fig. 4 shows the long-term spectrum common to both target and interference.

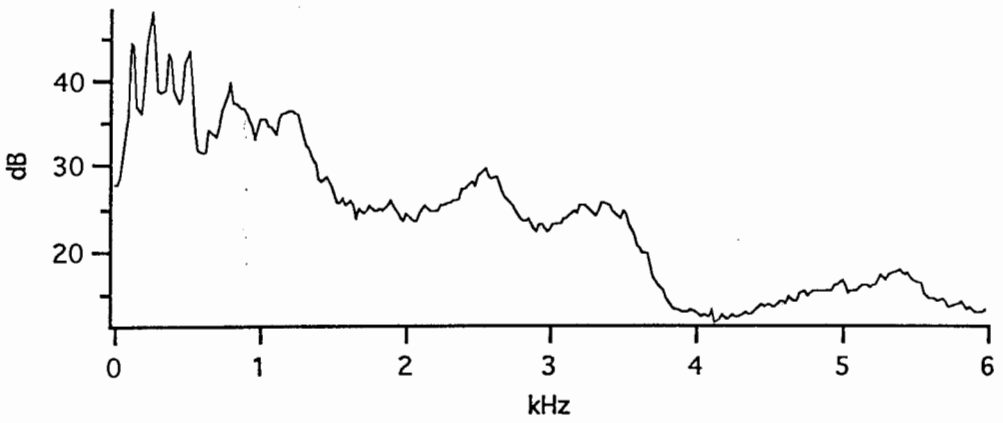

Fig. 4. Long-term spectrum of speech in database.

Target and interference were each voiced for 56% of all frames, and together for 41% of all frames. Fig. 5 shows the F0 histogram common to both target and interference speech. The distribution is relatively narrow, as reflected by the histogram of frame-by-frame F0 differences (Fig. 6). A large number of frames have a small F0 difference, making the task of separating them on the basis of F0 difference relatively difficult. Perceptual experiments suggest that a difference in F0 can be fully exploited to improve recognition as long as it is larger than 3 %. Such is the case of 75% of the frames for which both target and interference are voiced.

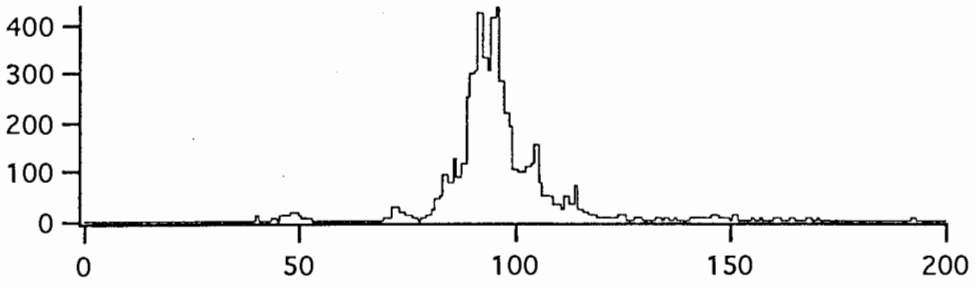

Fig. 5. Histogram of fundamental periods in database.

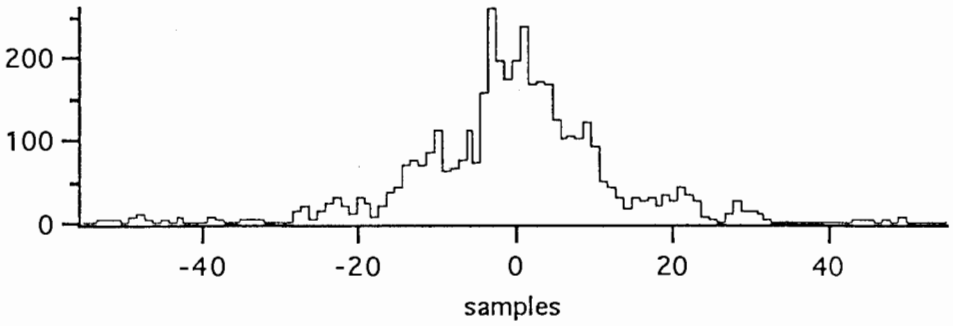

Fig. 6. Histogram of differences between F0s of target and interference.

2.4. F0 estimation

Pitch estimation used a standard AMDF (Average Magnitude Difference Function) algorithm. Details can be found in de Cheveigné (1993). In brief, the algorithm searches for an absolute minimum in the output magnitude of a comb filter (averaged over a time window). The period is taken to be the value of the lag (delay) parameter at the minimum: it is at this value that the input signal has been most effectively cancelled. The method is therefore appropriate for estimating F0 for cancellation purposes. The definition of the AMDF is:

$$AMDF(i,k) = \sum_{m=-N/2}^{N/2-1} |s_{i+m} - s_{i+m+k}|$$

where i is the analysis index and l the lag. In order to amplitude-normalize the function, eliminate the zero at zero lag, and attenuate spurious dips at short lags, the value at each lag was divided by the mean of values for shorter lags:

$$AMDF'(i,k) = AMDF(i,k) \Big/ (1/k) \sum_{m=1}^{k} AMDF(i,m)$$

The depth of the period dip in this function is the basis for a "periodicity measure" defined as:

$$\mathrm{PM} = -\log_2(AMDF'(T))$$

This measure gives an indication of the reliability of period estimation; it is large (2 to 6) where the speech signal is voiced and steady-state, and small during unvoiced parts and transitions.

The definition of lag used here corresponds to positive shifts of the signal index (the same convention is used later on for comb-filtering). F0 estimates are time-aligned to correspond with the middle of the analysis window.

To increase the resolution of F0 estimation, the signal was upsampled 4 times by linear interpolation, and low-pass filtered by convolution with a 1 ms square window. Integration window size N was 1600 samples (33.3 ms). The search range for the fundamental period corresponded to an F0 range of 60 to 300 Hz. Period estimates were produced at a frame rate of 2.5 ms, and expressed in terms of samples of the original sampling rate (12 kHz). All other processing was performed at the original sampling rate.

When the periodicity measure fell below 1.0 the speech was considered unvoiced and the period estimate was gated to zero. No other smoothing or error correction processing was used. Estimates were obtained from speech before mixing: no attempt was made to obtain them from mixed speech (see de Cheveigné 1993 for a discussion of this problem).

2.5. The time-domain cancellation comb-filter

The comb filter was implemented as a simple subtraction on the time-domain signal:

$$s'_n = (s_n - s_{n+l})/2$$

The lag parameter l was typically controlled by the F0 of the interfering voice. In the case of fractional F0 estimates, adjacent signal samples were interpolated. During non-voiced portions no filtering was applied; the onset and offset of filtering were smoothed by applying a raised-cosine ramp with a duration of 4.2 ms.

2.6. Filtering prior to mixing

The steps of addition and comb-filtering are both linear, and can therefore be swapped. Target and interference were therefore comb-filtered before mixing, with the same filter. This allows us to investigate separately the two factors that limit the effectiveness of voice cancellation: interference cancellation residue, and spectral distortion of the target.

2.7. Significance level of results

Formal significance tests were not done. However based on the criteria of the McNemar test (Gillick and Cox, 1989), one can give an *upper* limit of significance of individual differences. Individual differences of 5 words (5 %) or less fail to meet the 5 % significance criterion. Differences of 6 words or more may or may not meet this criterion, according to how the errors are distributed. Given the distribution trends observed, the main results (Fig. 16, 17) at least are statistically significant.

3. Factors that limit the effectiveness of voice cancellation

3.1. Interference cancellation residue

The residue is whatever is left over from the interference signal after filtering. If the interference were perfectly periodic, this residue would be zero. In practice, periodicity is rather imperfect, even within voiced sections, and a considerable amount of energy belonging to the interference is still present at the output. The input/output characteristics are detailed in

Fig. 7 for the interference signal (period equal to filter lag), and the target (period unrelated to filter lag).

The interference is attenuated most in the lower frequency region. The target is also slightly affected, which is understandable because target and interference have F0s that are close. Filtering gives the target a 6-7dB advantage within the 0-1.3 KHz range where most of the energy is concentrated (Fig. 4). This advantage tapers off to 2 dB at higher frequencies. During clean voiced portions of the interference the rejection ratio may locally be much greater. The interference residue is the inevitable consequence of the imperfect periodicity of speech.

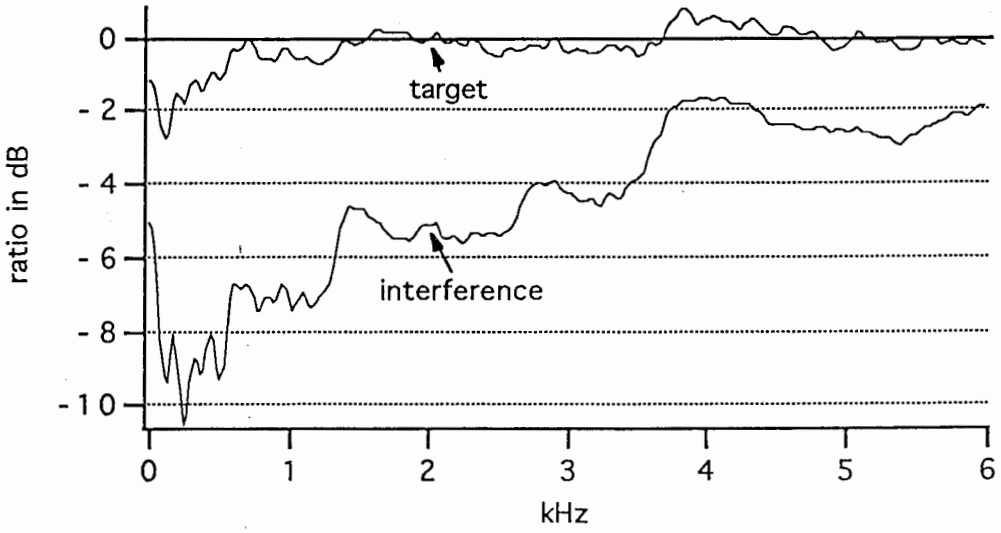

Fig. 7. Long-term spectrum rejection ratio of a comb filter tuned to the period of the interfering speech.

3.2. Spectral distortion of target.

The cancellation process inevitably also affects the target. Target components that happen to fall within the harmonic series of the interference are canceled, others may be attenuated. In the extreme case when the F0 of the target is equal to that of the interference or a multiple, the target is eliminated together with the interference. In the general case, the distortion can be described as an interference (moiré) pattern between the harmonic series of the target and the comb-filter frequency response.

The transfer function of the comb filter has an infinity of zeros equally spaced at multiples of the inverse of the lag parameter:

$$|H(f)| = |\sin(2\pi fL)|$$

Interaction with the target line spectrum is best illustrated by supposing that the target has a flat envelope (Fig. 8). The result of the interaction is illustrated in Fig. 9.

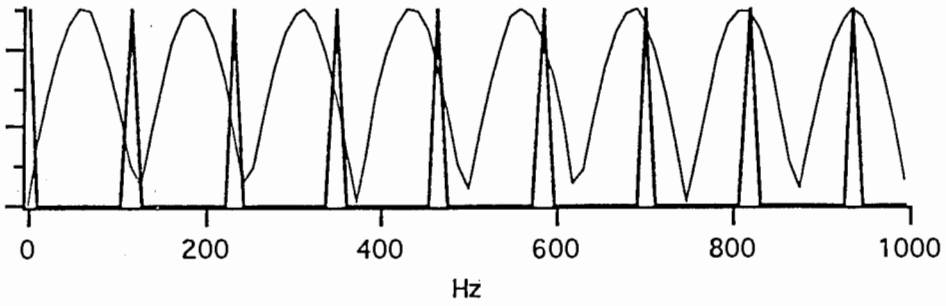

Fig. 8. Linear magnitude spectrum of target and transfer function of comb filter. Period of target and lag of filter are 102 and 96 samples respectively.

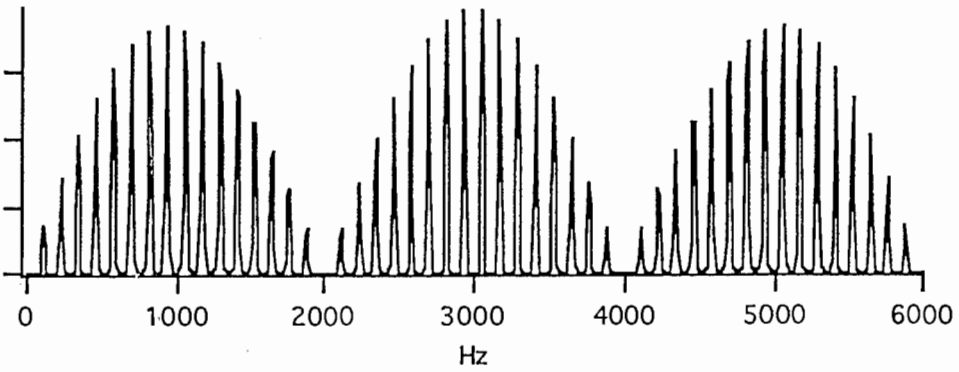

Fig. 9. Linear magnitude spectrum of output of comb filter.

The output spectrum is the input spectrum multiplied by a function:

$$|H(f)| = |\sin(2\pi f|T - L|)|$$

that is identical in shape to the transfer function of filter of lag L' = |T-L|. This transfer function has zeros at multiples of 1/|T-L|. The spectral distortion of the target can thus be described as the effect of comb-filtering at a "difference lag" that depends on fundamental periods of both target and masker.

The distortion can also be represented in the cepstral domain, where multiplication is represented by addition. The cepstrum of a cancellation-type comb filter of impulse response

$$h(n) = \delta(n) - \delta(n - N)$$

is (Rabiner and Schafer 1978):

$$p(n) = -(1/2)\sum_{r=1}^{\infty} \frac{\delta(n - rN)}{r}$$

The cepstrum representing the effect of comb-filtering a periodic signal with a lag that differs by 6 samples is plotted in Fig. 10. It shows peaks at multiples of the difference lag.

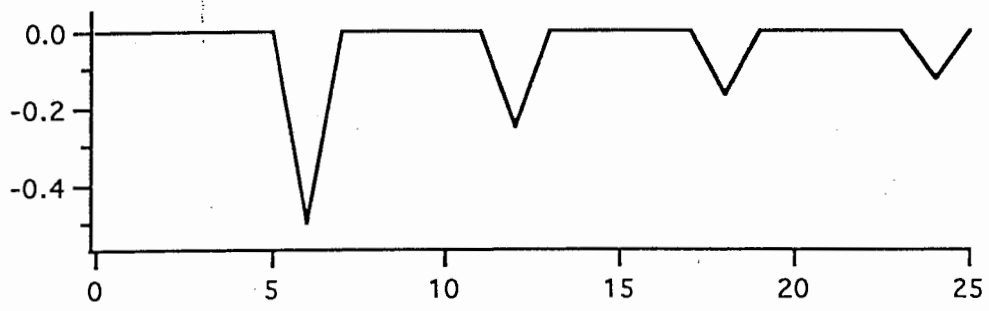

Fig. 10. Short-time portion of cepstrum representing the effect of comb-filtering a periodic signal.

This analysis assumes that the target is perfectly periodic, which is real speech is not. In particular, the shape of the spectral distortion is very sensitive to the width of the components of the spectrum. Simulation with real speech shows that actual distortion can depart considerably from this description.

4. Results

4.1. Recognition rate as a function of SNR.

The effect of adding interference to the target speech is shown in Fig. 14. With no interference (SNR = ∞) the rate is 100% which is as expected, as the task of matching words to templates belonging to the same set is trivial. Likewise when there is no signal (SNR = -∞) the recognizer is guaranteed to fail and the rate is thus 0%. At intermediate SNR the rates are relatively low, which is natural since interference signals belong to the same set as the target and are in therefore in direct competition. At an SNR of 0 dB the rate is less than 40 %. These rates constitute a baseline from which eventual improvements can be measured.

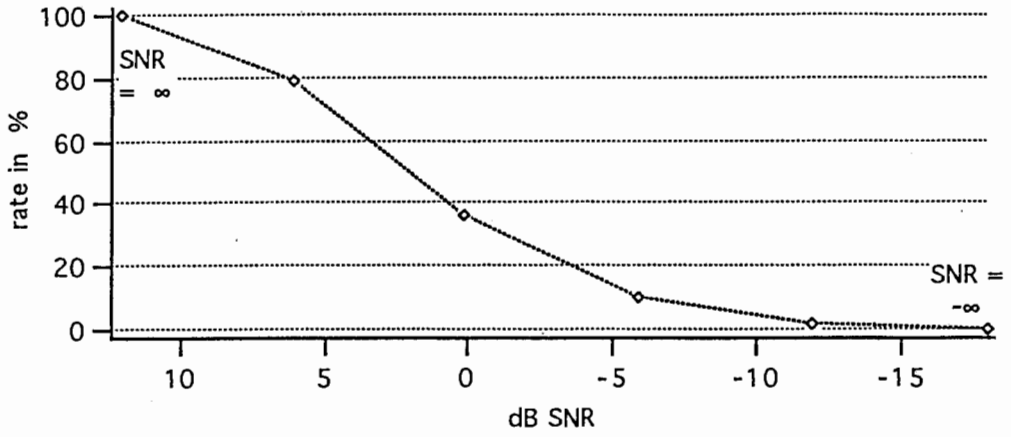

Fig 14. Recognition rate as a function of SNR for target mixed with interference.

4.2. Effect of the interference residue.

Here, the interference was canceled by a comb filter tuned to its period before mixing with the target. The target was thus unaffected by spectral distortion. This allows us to assess the effects of interference cancellation residue by itself. At each SNR level the residue was added with the same weight as in the uncanceled case (actual SNR was therefore higher than nominal SNR).

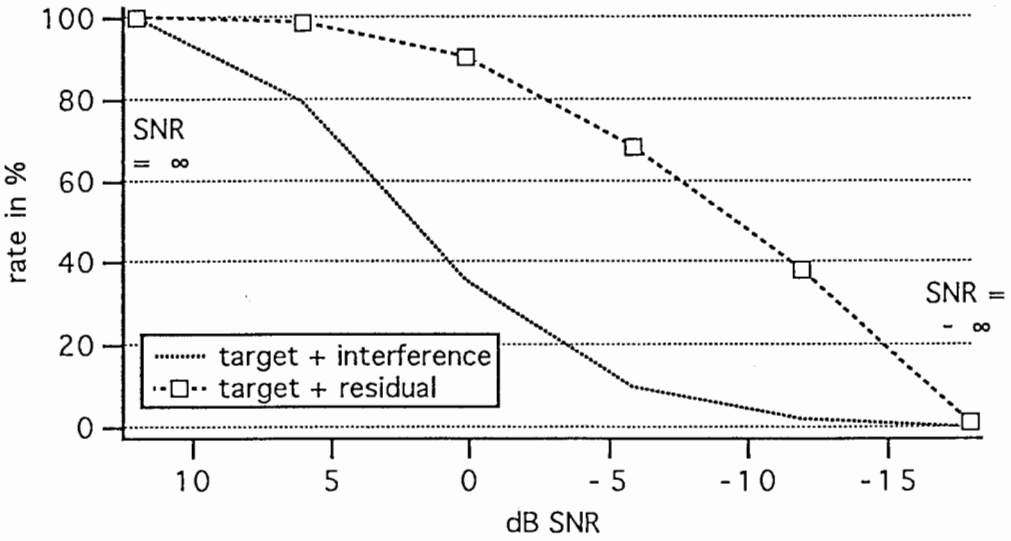

Fig. 15. Recognition rate as a function of SNR of target speech mixed with interference cancellation residue.

The distance from the 100 % line reflects the effect of the interference residue on recognition. It is of course greater at low SNR. The distance from the lower dotted line shows the effect of filtering the interference, about equivalent to reducing its level by 12 dB. This is more than one

would expect based on Fig. 7. That we observe a greater difference here is probably due to the fact that, in addition to reducing the amplitude of the interference, comb filtering also distorts it and makes it less apt to compete with the target for recognition.

4.3. Effect of spectral distortion of the target

The experiment was repeated applying comb filtering to both components. Comparison with the previous case allows us to measure of the effect of spectral distortion.

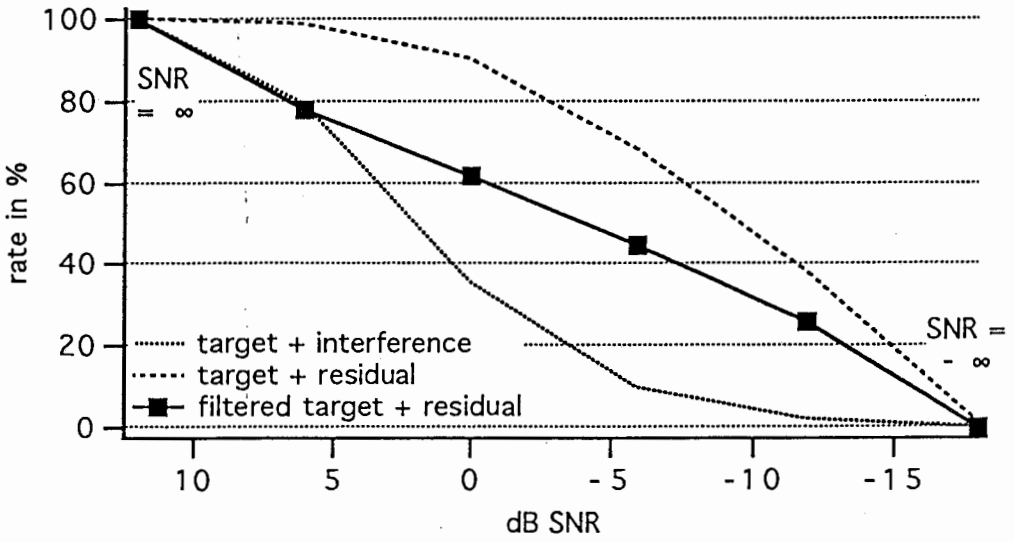

Fig. 16. Recognition rate as a function of SNR for comb-filtered mixed speech (filtered target + residue of interference cancellation).

Compared with no filtering, comb-filtering allows a clear increase in recognition rate at low SNR. At high SNR this increase is smaller, as the distortion introduced by the filter overcomes the benefit of interference reduction. At infinite SNR, spectral distortion does not affect the recognition rate, but this is most certainly due to a ceiling effect.

Spectral distortion has its greatest impact at high SNR, where it eliminates much of the benefit of filtering. This is a major problem for the design of a practical system, as such a system would probably only be of use when its performance is relatively reliable, ie at high SNR. Unfortunately the penalty of spectral distortion then outweighs any benefit of noise reduction.

4.4. Reference template adjustment

Spectral distortion disturbs recognition because the distorted targets don't match the templates so well. This suggests a scheme for reducing the effects of the distortion: distort the templates too, before matching. For this

we must know precisely the spectral distortion that affects the target. The analysis of section 3.2 suggests a way to estimate this distortion, knowing the fundamental frequencies of both target and interference.

Template adjustment was implemented in the frequency domain by multiplying the reference template spectral features by the transfer function of a comb filter tuned to the difference between the periods of target and interference. The result is presented in Fig. 17.

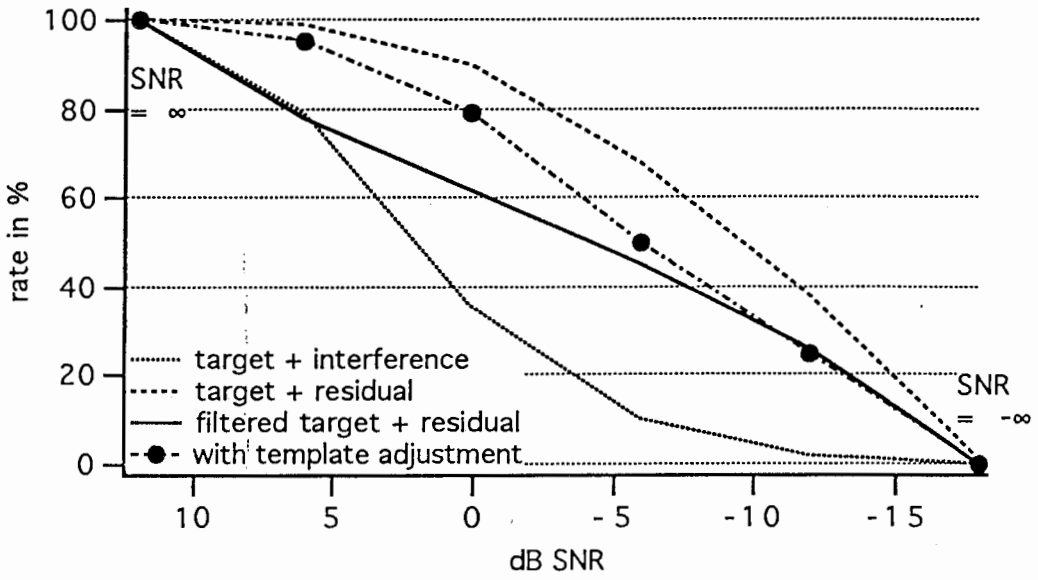

Fig. 17. Recognition rate as a function of SNR. The line with markers is for comb-filtered mixed speech with reference template adjustment.

At low SNR there is little benefit, but at high SNR template adjustment allows us to regain much of what was lost to spectral distortion. If this result is confirmed for more realistic tasks and more mainstream recognition techniques, the scheme might be of practical use for noise reduction in speech recognition systems.

4.5. Frequency vs time-domain processing.

Most other published schemes for voice separation work in the frequency domain. Time-domain processing has several advantages: the processing filter can be short, and it can adapt quickly to changes in the F0 of the interference. On the other hand frequency-domain filtering is more "predictable" in terms of spectral distortion. Fig. 18 shows the rates obtained when the time-domain cancellation filter is replace by the "equivalent" frequency domain comb filter, based on the mean period value within the analysis frame.

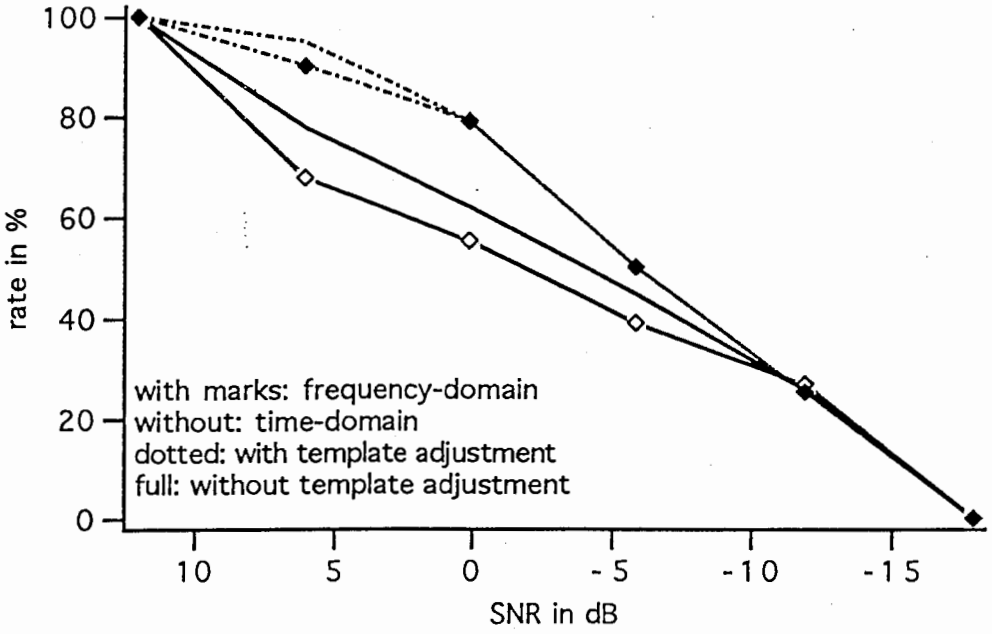

Fig. 18. Recognition rates as a function of SNR. Lines with marks are for frequency-domain cancellation. Lines without marks are for time-domain cancellation (same as data in Fig 17). Upper lines are with template adjustment, lower lines are without.

The differences in rate are too small for reliable interpretation. The most one can say is that they do not contradict our interpretation that time domain processing may be more effective for cancellation, but that spectral distortion produced by frequency-domain processing is more accurately represented in the matched distortion applied to reference templates.

A fair comparison of time-domain and frequency-domain processing would require implementing the more sophisticated techniques that have been proposed in the literature (for example Parsons 1976).

4.6. Enhancement vs cancellation.

In principle, one can just as well enhance a harmonic target as cancel harmonic interference. Each strategy has advantages and disadvantages, as listed below:

Cancellation:

- Works whatever the target (vowels, consonants, etc.) but the interference must be harmonic.
- F0 of interference can be better estimated when SNR is low, so cancellation is easier in this case.
- Cancellation causes spectral distortion of the target.
- Cancellation can be implemented with a filter with a very short impulse response.

Enhancement:

- Works whatever the interference, but the target must be harmonic (only voiced parts can be enhanced).
- F0 of target can be better estimated when SNR is high, so enhancement is easier. But speech separation is less necessary in this case.
- Enhancement does not cause spectral distortion (if the target is perfectly harmonic).
- Effective enhancement requires a filter with a long impulse response (maybe impractical because speech in non-stationary).

Some of these arguments depend on the harmonicity and stationarity of target and/or interference, and it is difficult to predict how they apply to real speech. I therefore compared the two using our speech recognition paradigm. Enhancement can for example be implemented with a comb filter defined by the following impulse response:

$$h(t) = (1/N)\sum_{k=0}^{N-1}\delta(t - kL)$$

This consists of N "prongs" equally spaced at intervals of the lag parameter L, which is adjusted to the period of the signal to be enhanced. In theory, the ratio of enhancement of the target is equal to the number of prongs. A high ratio therefore requires a long impulse response, which may then be less effective due to the non-stationarity of the target. The practical enhancement ratio (output SNR for a 0 dB input SNR) for the database is plotted in Fig. 19.

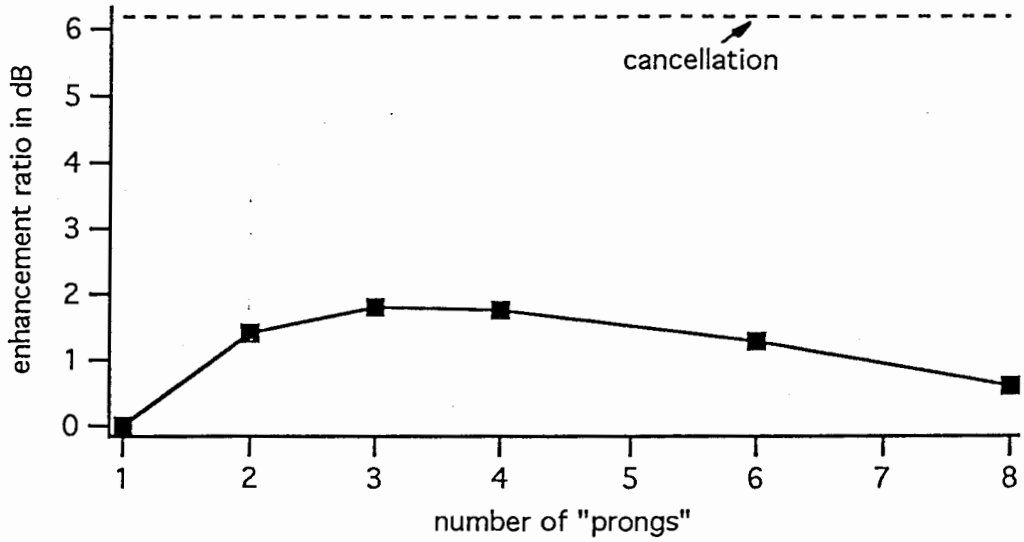

Fig. 19. Enhancement ratio as a function of the number of "prongs" in the impulse response of the enhancement filter. The dotted line shows the ratio attained by a cancellation filter.

Enhancement is at its maximum for a 3-pronged filter. Adding prongs makes the filter less effective. The ratio is far from its theoretical value (6 dB for 2 prongs), and from the ratio attainable using cancellation. Adding prongs also increases spectral distortion (which should in theory be zero for enhancement). This can be quantified by defining a measure of spectral distortion, based on the spectral feature vectors before (f) and after (f') processing:

$$d = \left(\sum_{k=1}^{16} (f_k - f_k')^2 / \sum_{k=1}^{16} f_k^2 \right)^{1/2}$$

The spectral distortion measure is plotted in Fig. 20. It remains less than that caused by cancelation until the filter is 8 prongs in length.

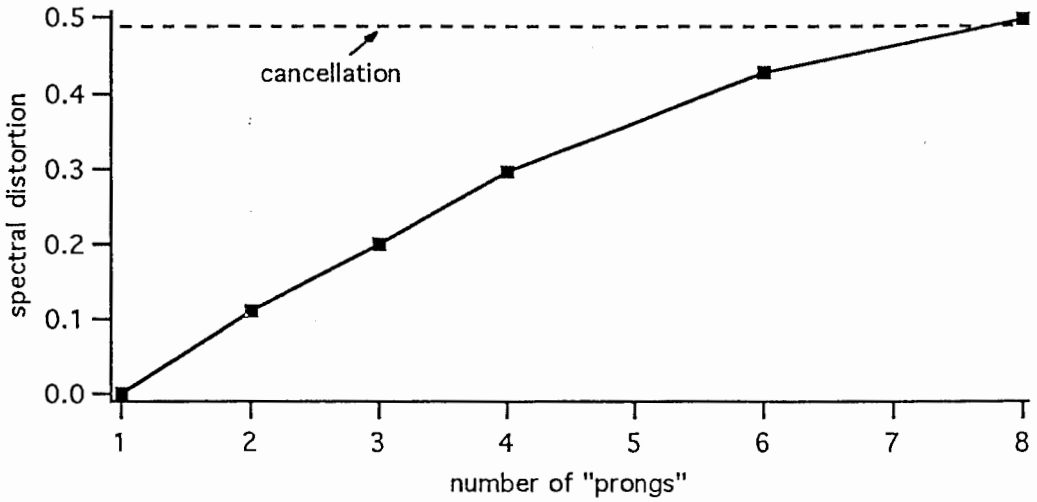

Fig. 20. Spectral distortion as a function of the number of "prongs" in the impulse response of the enhancement filter. The dotted line shows the spectral distortion produced by a cancellation filter.

The recognition rates are plotted in Fig. 21. At most SNR levels a 3-prong filter gives the best rates, which remain far below those attained by the cancellation filter (with template adjustment). Enhancement is clearly less effective than cancellation for this task.

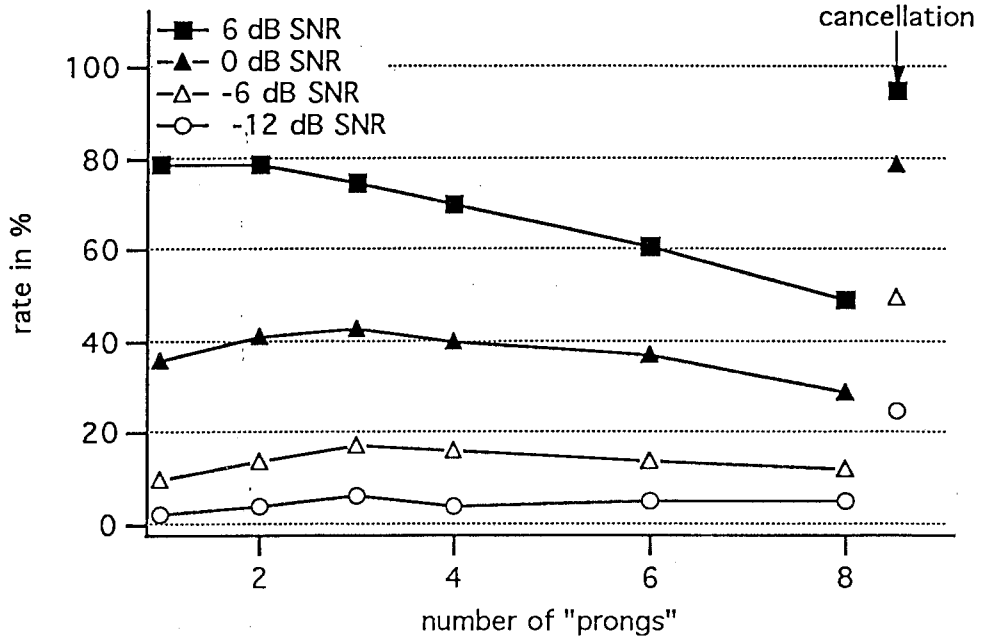

Fig. 21. Recognition rate as a function of the number of "prongs" in the impulse response of the enhancement filter. The rates for N=1 are the same as for no filtering.

Cascaded cancellation and enhancement filters provide the best rejection ratio, and a distortion ratio below that of cancellation alone, but the recognition rates are nevertheless less good than for cancellation alone (not shown).

4.7. The effect of F0 mistuning.

It is interesting to know how the accuracy of F0 estimation affects the effectiveness of filtering. To investigate this question, I systematically mistuned the F0 estimates used for comb filtering. The effect on rejection ratio is shown in Fig 22. Mistuning has little effect on spectral distortion.

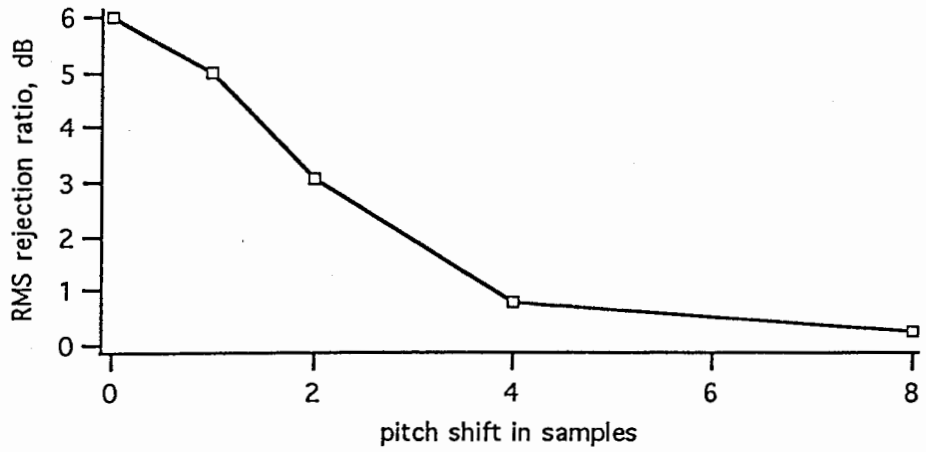

Fig. 22. RMS rejection ratio as a function of F0 mistuning.

Recognition rate as a function of mistuning is shown in Fig. 23. The rates level off after 4 samples (approximately 4 %). This means for example that, to cancel interference effectively, a system must estimate F0 with an accuracy better than 4 %. Also shown is the rate when the resolution of F0 estimation is limited to 1 sample, rather than the 1/4 sample resolution obtained by upsampling.

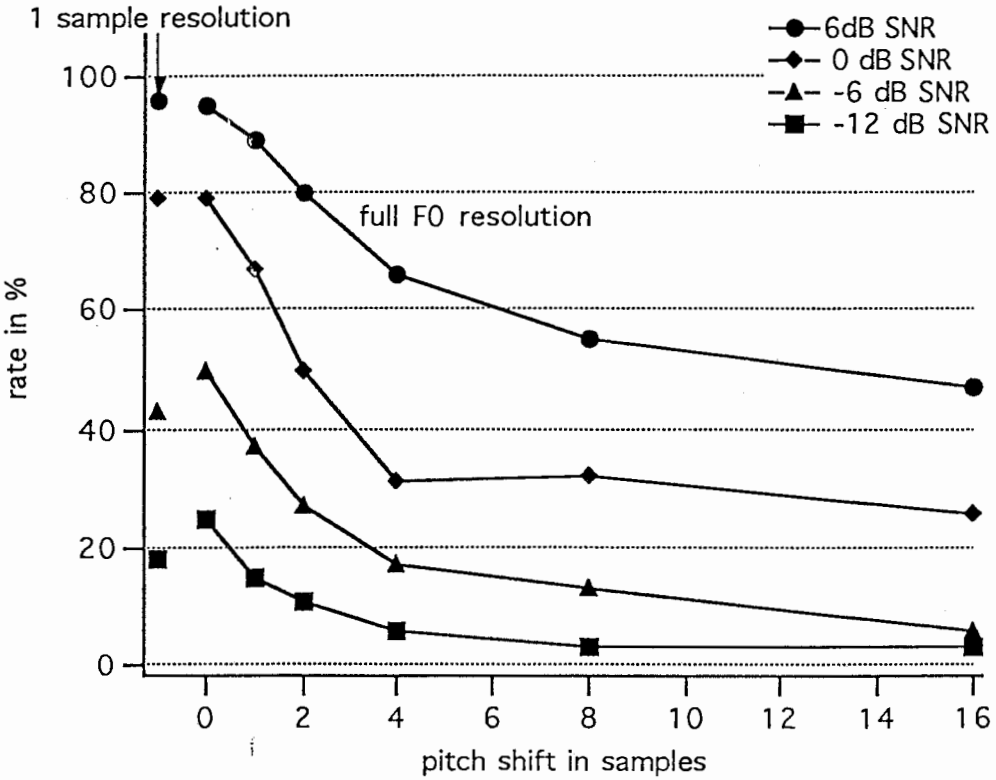

Fig. 23. Recognition rate as a function of F0 mistuning, for different signal-to-noise ratios.

It is interesting to compare the tuning of the cancellation filter with the tuning of rates measured in psychoacoustic "double-vowel" experiments. Fig. 24 shows the rates of recognition of synthetic vowels by human subjects, obtained by Culling and Darwin (1993). The rates are for individual vowels in a mixture, as calculated by taking the square root of the published "both-correct" rates. These experiments and theirs are of course for the most part incomparable, but it is interesting to note that the levelling off of recognition rate as a function of mistuning occurs in a similar region.

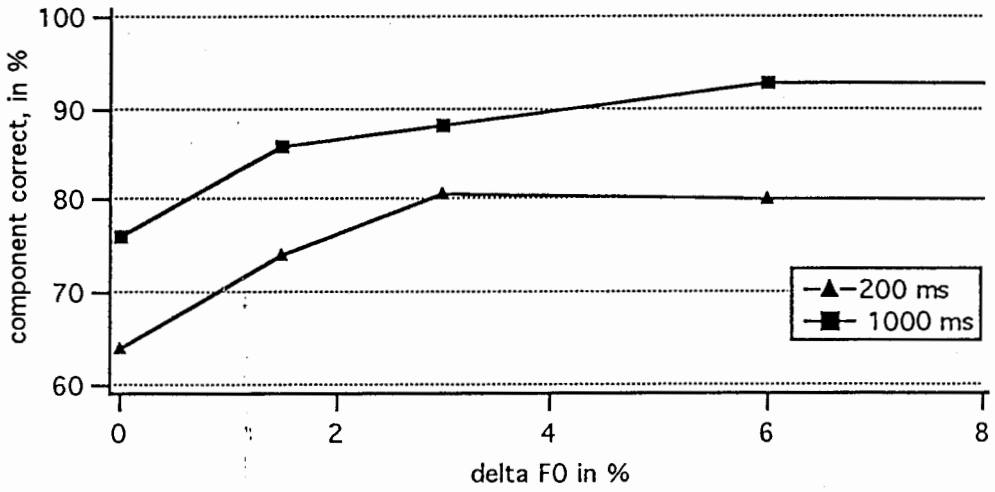

Fig. 24. Rate of recognition of synthetic vowels in a pair, by human subjects, as a function of difference in F0.

4.7. Cepstrum vs spectrum.

The experiment was repeated using a 16-point cepstrum representation with Tohkura-weighting, instead of the 16-point linear magnitude spectrum representation.

The cepstrum representation appeared to be less sensitive to interference, as evident in Fig 25. The effect of filtering is also smaller, equivalent to 2 to 3 dB increase in SNR, as opposed to 9 dB for the linear spectrum representation. The rates obtained after filtering are thus less good. Template adjustment was implemented in the spectral domain (before calculation of the cepstrum) and also in the cepstral domain as either a filter (subtracting the ideal comb-filter cepstrum) or a lifter (setting to zero, in the target and reference cepstra, the coefficients for which the comb-filter cepstrum is non-zero, fig 26). The effect of template adjustment was very small (not shown).

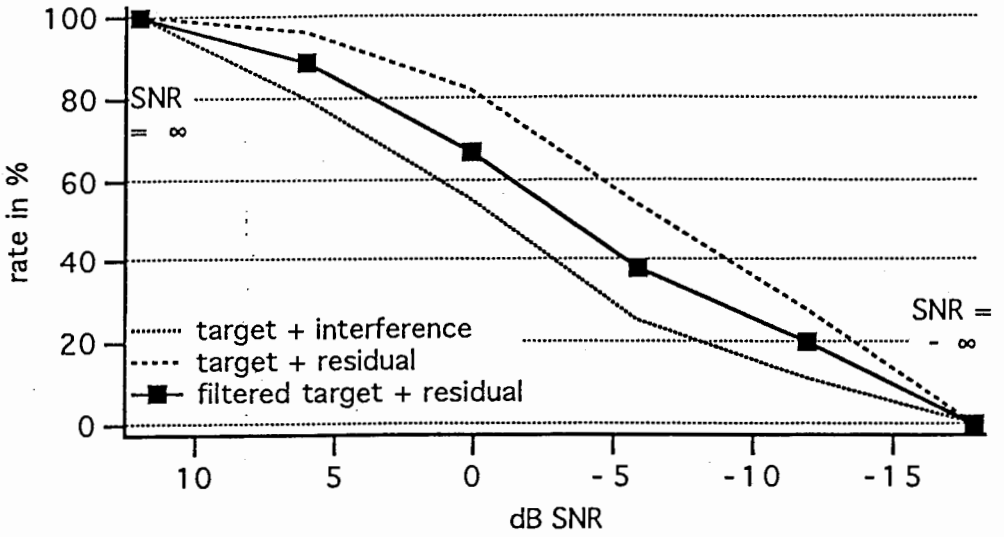

Fig. 25. Recognition rate using cepstral coefficients, as a function of signal to noise ratio.

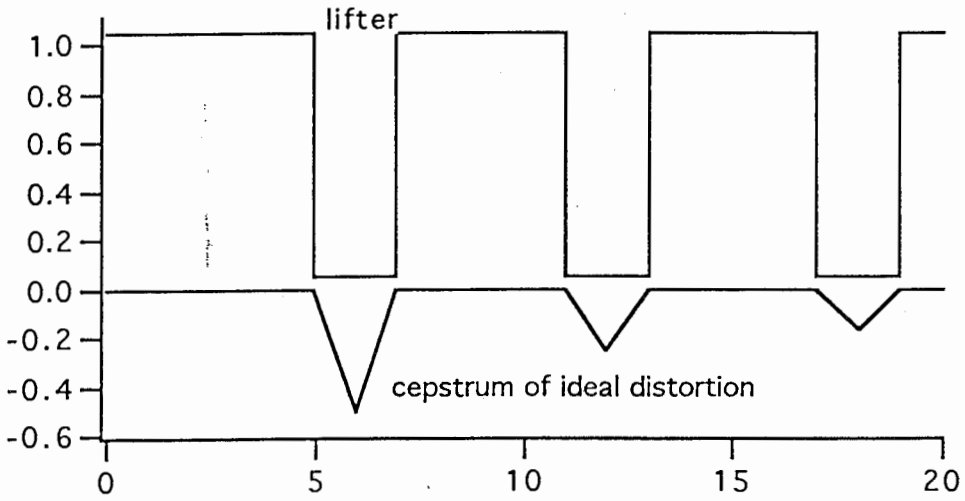

Fig. 26. Top: lifter used to mask distorted cepstral portion of both target and reference. Bottom: cepstrum representing ideal distortion.

Similar results were found when log spectrum coefficients were used. The discrepancy with the linear representation can be interpreted by saying that the linear representation puts a strong weight on the portions of the pattern that have a strong amplitude. Since amplitude correlates with voicing, it gives a strong weight to portions where the harmonic separation schemes are effective. On the other hand the cepstrum and log-spectrum representations give an equal weight to low and high amplitude portions, and thus dilute the effects of harmonic separation. Which representation is

most appropriate in applications that require speech separation is a question that needs to be investigated.

4.8. Some Things it Would be Nice to Try.

1) A better database. This one is too small. Reference words and target words are identical, so the recognizer may use cues that are particular and unreliable. The F0s have a too narrow distribution.

2) Data representation based on cepstrum features weighted by energy (or some power of energy, for example 1/2). This would allow us to test our interpretation of the discrepancy between linear spectrum and cepstrum coefficients, and possibly combine the advantages of both.

3) Low-pass filtering. Fig. 7 suggests that cancellation is more effective at low frequency, so restricting the features to that portion may prove effective.

4) Formal tests of the significance of results.

5. Conclusions

The experiments brought relatively clear answers to the questions we formulated, but may not be clear at this stage which are pertinent for perception models, and which for speech processing.

The conclusions one can draw concerning perceptual models of harmonic sound separation are:

1) Time-domain processing is effective for interference reduction, even given signals with imperfect periodicity such as speech. Before extending this conclusion to a neural model such as proposed in 1.1, one must however investigate in detail how processing effectiveness might be affected by differences in signal representation (neural vs linear) and processing.

2) In a task where harmonic enhancement and harmonic cancellation were both a priori possible, enhancement was much less effective than cancellation. It is likely that this conclusion is valid also for auditory processing in similar conditions.

3) Knowledge of the distortion caused by filtering can be used to reduce the effects of distortion on pattern matching. Another way of formulating the process of filtering-plus-template adjustment is to say that the system puts zero weight on the portions of the spectrum that contain information that it cannot attribute with certainty to the target. Such a strategy might also be at work in the auditory system. Non-uniform weighting of information can be applied along other dimensions as well, such as time. This suggests an explanation of CMR effects: knowledge of the interference,

gained from monitoring the amplitude of off-signal channels, is used to apply a non-uniform weight to on-signal channels. In effect, the auditory system listens for the signal in valleys of the masker.

The conclusions that one can draw concerning processing schemes for voice separation are:

1) Time-domain comb-filtering is effective for reducing the effects of an interfering voice on speech recognition. Some evidence was found to the effect that it is better than frequency-domain processing because it can cope more easily with non-stationarity, but comparisons must be made with more sophisticated schemes before firm conclusions are drawn. One can argue that time-domain processing makes full use of the harmonic structure: to go beyond requires other assumptions, such as continuity in the timbre domain, etc..

2) The effects of spectral distortion can be reduced by adjusting the reference features to match the distortion of the target features. This boosts performance in the high-SNR region important for applications.

3) Less clear results were obtained when cepstrum features were used. This question requires further examination, as cepstrum or log spectrum features are more often used for speech recognition than linear spectrum features.

4) Harmonic cancellation requires that the F0 of the interference be estimated with an accuracy of at least 4%. The template adjustment scheme requires knowledge of the F0 of both target and interference.

6. Bibliography

de Cheveigné, A. (1993), "Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing", J. Ac. Soc. Am. 93, 3271-3290.

Culling, J.F., and Darwin, C.J. (1993), "Perceptual separation of simultaneous vowels: within and accross-formant grouping by F0", J. Ac. Soc. Am. 93, 3454-3467.

Gillick, L., and Cox, S.J., "Some statistical issues in the comparison of speech recognition algorithms", IEEE ICASSP, 532-535.

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

Rabiner, L.R., Schafer, R.W. (1978) "Digital processing of speech signals", Prentice-Hall, Englewood Cliffs, NJ.

Tohkura, Y. (1987) "A weighted cepstral distance measure for speech recognition", IEEE Trans. ASSP, 35, 1414-1422.