Internal Use Only（非公開）

TR－SLT－0047

Noise robust sub-band ASR using a HMM/BN framework

Etienne Denoual

August 29, 2003

This document will present the research work I conducted in ATR for five months from April to August 2003. Many approaches have been proposed in the past to deal with noise in robust speech recognition, mainly feature-based and model-based. Here we consider the use of a hybrid HMM/BN framework to increase ASR robustness, by computing the posterior likelihood of noise in each frame's sub bands, and then marginalizing corrupted channels.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２ TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

Table of contents

Table of Figures

Table of Tables

Acknowledgments

I would like to thank Dr.Satoshi Nakamura (ATR-SLT Dept.1 Head), Dr.Mitsunori Mizumachi (ATR-SLT researcher) and of course Dr.Konstantin Markov (ATR-SLT Senior researcher) my supervisor, for their scientific and technical support as well as for their warm welcome in ATR and in Japan .

Thank you to all members of SLT, especially Hiroko Kawa and Kahoru Ohtsuki (Planning section ATR-SLT) for their every day support and help on many things, and Shigeki Matsuda, Toshiki Endo and Toshiharu Horiuchi (ATR-SLT researchers) for their friendliness and kindness.

Than you to Yukiko Ishikawa and Makiko Tatsumi (SHIEN group ATR) for making our life in Japan so much easier, day-by-day.

Thank you to Laurent Besacier (CLIPS-France "Maitre de conferece") for his support from France.

Thank you to Mathieu, Yves and Eddy for making my stay in Japan so pleasant.

And finally thank you to Nicolas, Christian, Amori, Chiyako and Yuichiro for making me discover another side of Japan.

Introduction

Over the years, research and performance in the field of speech recognition has greatly evolved and improved : new systems are developed that achieve better and better results and spawn numerous user-friendly applications in mobile telecommunication systems, human-machine interfaces, or real time word processing software . However one point remains to be greatly improved : the robustness of such recognition .

Excellent recognition rates are achieved for what we refer as "clean"/ideal conditions, that is "laboratory" conditions of clean data, the absence of surrounding noise and a stable environment . In everyday life these conditions are never met : there is always noise (other interfering dialogues, reverberation of the speaker's voice, natural background noise), and in spontaneous speech speakers tend to introduce hesitations, or different stress patterns to what they say ; those elements should not be considered as useful information by the ASR system .

The main point to focus on in modern speech recognition nowadays is therefore that of robustness . There are two main paths to follow for the researcher : focus on the recognition itself (pre-processing and processing), that is noise adaptation, compensation, improving acoustic models, or on the post-processing of this part, focusing on language models to re-estimate the reliability of the results , and using more sophisticated models or confidence measures.

During my internship in ATR I have explored the first path, using a sub-band approach in an attempt to detect noisy bands and therefore eliminate noisy data, i.e. corrupted feature vectors. Sub-band techniques have already been widely used for speech recognition [] . These techniques aim at extracting useful information from different bands in order to produce "good" information to the recognizer, which implies various ways selecting, combining, correcting. A pre-requisite for these methods is a good detection of noisy speech. For this matter, here I will investigate the use of a hybrid Hidden Markov Model (HMM) / Bayesian Network (BN) framework. The advantage of this framework is that while it may continue to benefit from existing and efficient HMM principles, we can use a Bayesian Network to characterize speech frames as they come, and orient recognition with this new knowledge by discarding noisy frames or using a noisy-trained model instead

I will first present further background on the research topic, then the objectives and approaches that were fixed at the beginning of the internship . Then at each step I will go into more detail concerning the methods that were developed, along with experimental results to validate/confirm such methods. Finally I will reflect on the overall achievement of my research compared to the objective chosen at the beginning and on personal assessments.

Noise Detection and Speech Recognition using a Hybrid HMM / Bayesian Network Framework in a sub-band approach

Objective and global approach

The objective here is obviously to get better recognition rates using our new hybrid HMM/BN sub-band system than with using standard full band features with HMM. In the process of building and testing such system we will go through the following steps :

Implementing Sub-band feature extraction

After designing and benchmarking a standard full-band baseline, it will be interesting to alter feature extraction and processing for a certain variety of band configurations. Such configurations and experimental setup is best described in [1] as we will see further

Producing a 5-band hybrid model and artificial testing data

The next step will then be to merge in certain sub-band configurations a clean-trained and a noisy-trained model, in order to detect noise according to our Bayesian relationship. For this matter it will also be interesting to produce artificially corrupted/noisy data, be it locally or widely.

Building & Benchmarking the hybrid system

The final step will then be to build a recognizer capable of taking into account our new framework: after this is done we can begin to refine the noise detection using artificially corrupted data, and then to benchmark our system and its several refinements using a variety of artificial, then standard "natural" test data. Here we will use the English AURORA 2.0 database (noisy TIDigits database)

Baseline system

The first step before implementing any new method was to get some baseline system running in order to improve its results. Through the whole process I used a specific software package designed to build hidden Markov models (HMM), called HTK for HMM Tool Kit. Such tool kit includes tools for generating and training HMMs given some specific training data, and recognizing tools to test these HMMs according to some test data.

The experimental setup used is as follows :

- 20 dimensional feature vectors including 10 Mel-scale Frequency Cepstrum Coefficients (MFCCs), and their 10 first order derivatives (commonly referred to as MFCC plus delta). For more information on MFCCs, please refer to appendix I.
- The acoustic model consists of 3 mixtures of gaussians per state models.
- The training data used was the clean training data from the AURORA 2.0 database as described more thoroughly in [6], and consists in 8440 utterances from the TIDigits database by 55 male and 55 female adults. These signals are filtered with the G.712 filter to consider the realistic frequency characteristics of terminals and equipment in the telecommunications area .
- The testing data is also form the AURORA 2.0 database. It consists of the two first subsets included (there are three in total). 4004 utterances from the TIDigits database are split into 4 subsets of 1001 utterances each. Recording of all speakers are present in the two subsets. One noise signal is added to each subset of 1001 utterances at the different SNRs of 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB. Also, the clean case without added noise is taken as seventh condition. Again, speech and noise are filtered with the G.712 filter characteristic before adding. Test set a includes suburban train, babble, car and exhibition hall noises, whereas test set b includes restaurant, street, airport and train station noises.

The results are shown in table 1. Those are word accuracies given in percent (%) according to the following computation :

$$WordAcc = \frac{N - S - I - D}{N}$$

where WordAcc is the word accuracy, N the total number of emitted words, S the number of substitution errors, I the number of insertion errors and D the number of deletion errors.

A				
	Subway	Babble	Car	Exhibition
Clean	97.97	97.88	97.79	97.78
20 dB	95.24	93.11	95.68	95.65
15 dB	89.50	88.12	88.58	88.40
10 dB	72.21	75.12	67.91	63.25
5 dB	40.84	50.39	33.55	30.64
0 dB	21.77	23.13	18.70	18.08
-5dB	14.09	10.94	12.08	11.76
Average	61.66	62.67	59.18	57.94

Test set B					
	Restaurant	Street	Airport	Station	Average
Clean	97.97	97.88	97.79	97.78	97.86
20 dB	91.53	95.13	91.14	93.18	92.83
15 dB	86.71	87.27	85.24	88.03	86.81
10 dB	74.12	66.51	73.99	73.16	71.94
5 dB	49.22	39.12	48.08	44.89	45.33
0 dB	23.15	22.37	23.41	20.24	22.29
-5dB	11.24	13.85	11.06	11.05	11.80
Average	61.99	60.30	61.53	61.19	

Table 1 : Baseline recognition results

A word on sub-band methods in ASR

The first step in most of the current automatic speech recognizers (ASR) is to convert the incoming speech signal into series of short-term vectors (feature vectors). Each vector represents a short segment of the signal (also called "frame" or "observation"). Each element of the vector describes some part of the information carried by the signal; for example each element of the feature vector represents energy of the speech signal in a given frequency range.

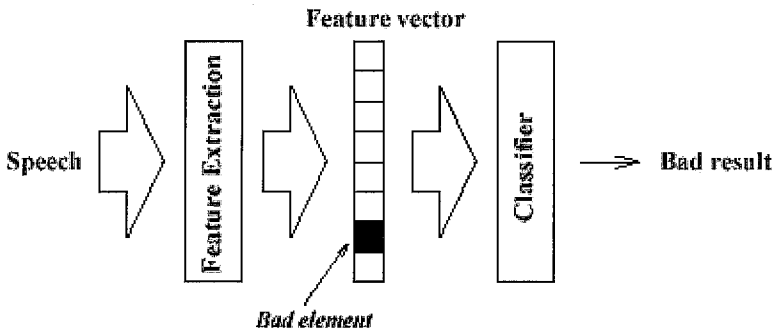

Figure 1 Conventional ASR

Consider the case when some of the elements of the feature vector contain corrupted or even misleading information, while the remaining ones are still uncorrupted. This can occur e.g. when the speech signal gets corrupted by selective noise. In the current mainstream ASR the entire feature vector is used as one entity and even a single corrupted spectral element can severely degrade the performance of the recognizer.

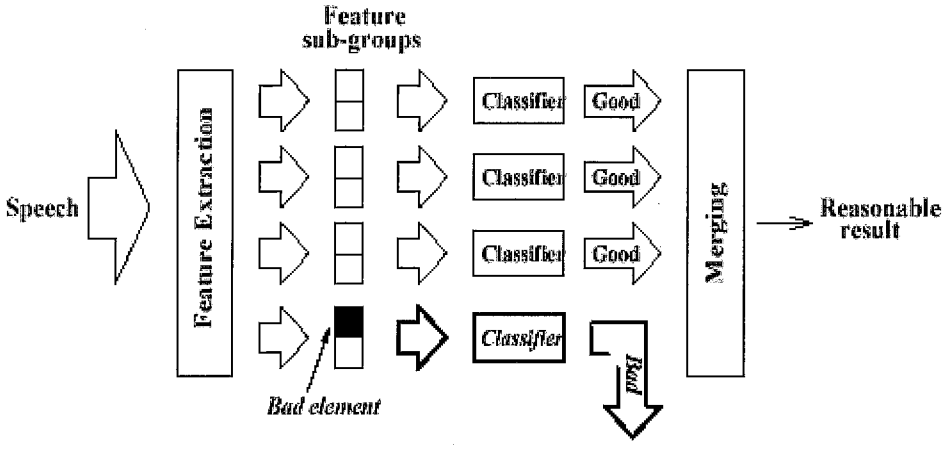

Figure 2 Sub band model

It is argued in [4] that human auditory perception works differently than the current ASR. More specifically, it is suggested that the linguistic message gets decoded independently in different frequency sub bands and the final decoding decision is based on merging the decisions from the sub-bands.

One interpretation is that as soon as any sub-band combination yields sufficient information, the information from the remaining (possibly corrupted) sub-bands does not have to be used for subsequent decoding of the message, and can therefore be marginalized. It is this approach that will be used in the following work.

Sub-band feature extraction

In this part we will alter the way features are extracted from raw data : commonly we use cepstral parameters called MFCCs for Mel-Frequency Cepstral Coefficients and calculated from the log filter bank amplitudes using discrete cosine transform (DCT). The mel-scale is logarithmic as argued in [8], empirical evidence has shown that the human ear resolves frequencies non-linearly, and that the use of such scale gives better recognition results.

Futher details about Filterbank analysis can be found in appendix I

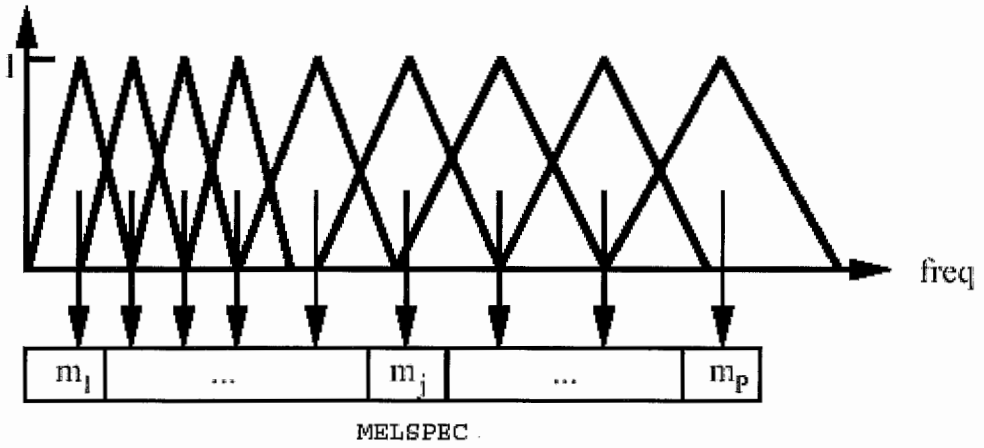

Figure 3 Mel-Scale Filter Bank

In a fullband approach MFCCs are calculated from the log filter bank amplitudes using the Discrete Cosine Transform :

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right)$$

where N is the number of filter bank channels. This produces the following acoustic vector (here we take our baseline example of 20 dimension vector – 10 ceps plus 10 delta) :

C_1	C_2	...	C_{10}	D_1	D_2	...	D_{10}

Figure 4 20-dimensional acoustic vector (MFCC plus delta) – full-band

Let's now consider sub-band feature extraction : we will here take the example of a 2band system. The first half of the filters will contribute to calculate the first series of MFCCs plus delta (band one), the other half will contribute to calculate the ones of band two, according to the following pattern :

$$c_{1i} = \sqrt{\frac{2}{N_1}} \sum_{j=1}^{N_1} m_j \cos\left(\frac{\pi i}{N_1}(j-0.5)\right) \text{ and } c_{2i} = \sqrt{\frac{2}{N_2}} \sum_{j=1}^{N_2} m_{j+N_1} \cos\left(\frac{\pi i}{N_2}(j-0.5)\right)$$

Where $N_1 = N_2 = \frac{N}{2}$ if N is even and $N_1 = \frac{N+1}{2}$, $N_2 = \frac{N-1}{2}$ if N is odds. This therefore outputs the following acoustic vector :

C_{11}	...	C_{15}	C_{21}	...	C_{25}	D_{11}	...	D_{15}	D_{21}	...	D_{25}

Figure 5 20-dimensional acoustic vector (MFCC plus delta) – 2bands case

I therefore modified and compiled the HTK sources (The feature extraction tool in HTK is called HCopy) in order to implement feature extraction for 2,3,5 and 7 band systems. Number of filter banks and desired MFCCs were specified in an external configuration file, and automatically computed.

Furthermore the aim was to process these new acoustic vectors using the streams feature in HTK, which allows one to split an acoustic vector into a separate number of data streams to be computed separately. I therefore built a tool to analyze a MFCC file (observation file), such as the one regrouping several dozens of feature vectors like the one above in the following manner :

Stream1 5MFCC+5Delta						**Stream2** 5MFCC+5Delta					
C_{11}	...	C_{15}	D_{11}	...	D_{15}	C_{21}	...	C_{25}	D_{21}	...	D_{25}

Figure 6 "swapped" 20-dimensional acoustic vector (MFCC plus delta) – 2bands case

This tool performed automatically, reading HTK headers and configuration file, reassigning filter banks to streams by itself. Along with this tool I had to create a batch processing means (this swap was to be performed on several thousand utterances as we have previously seen).

When this was done, all data both training and test was processed for the following experimental conditions:

Sub-band configurations	Filter banks	Filters/Sub-band	Mfcc/Sub band	All Mfcc (Delta)
2	30	15	7	14 (28)
3	30	10	5	15 (30)
5	30	6	3	15 (30)
7	28	4	2	14 (28)
1 (full band)	20	20	10	10 (20)

Table 2 : Sub band extraction experimental conditions

This concludes the phase of sub band extraction and data preparation.

Producing a 5-band hybrid model and artificial testing data

As in [1], we will consider a 5 bands approach that appears to be a better choice in terms of the balance between the noise localization and phonetic discrimination. However, to comfort this opinion it was interesting to build 2,3, and 7 bands systems to compare with.

The results are shown in the figure below : word accuracy here is shown as an average of the 8 different noise conditions from AURORA 2.0 given a common SNR.

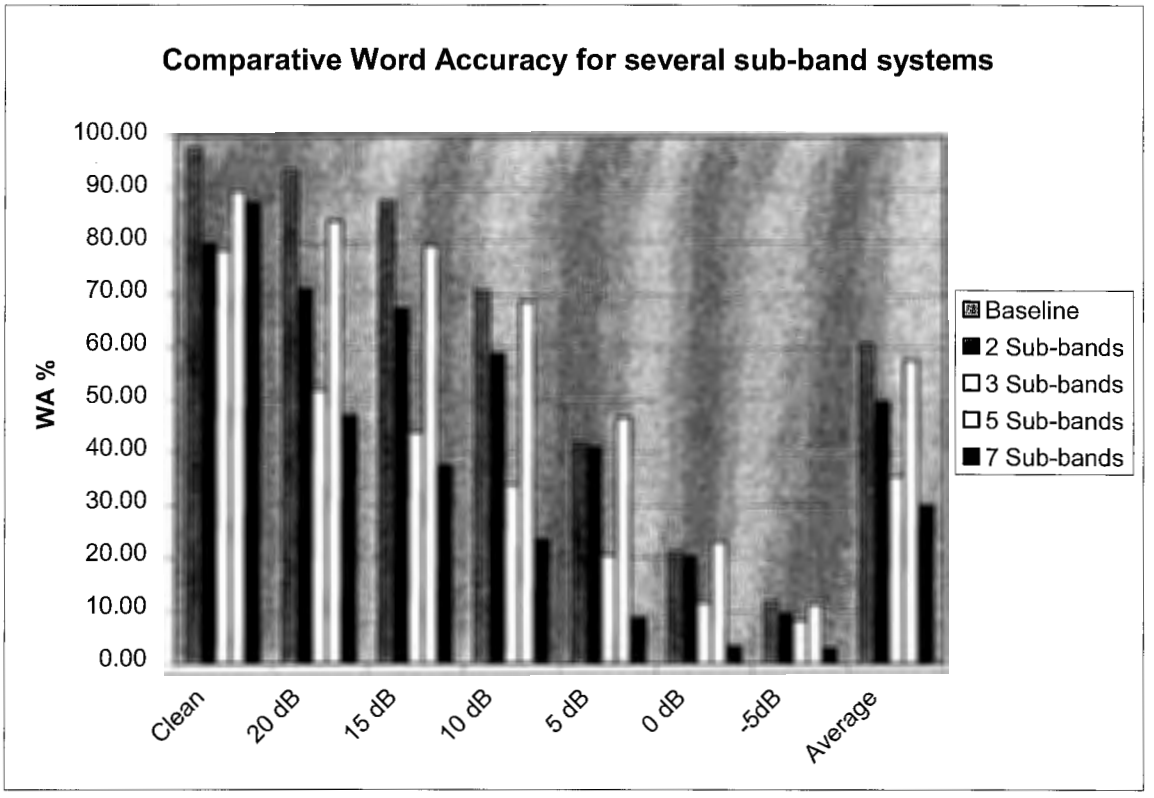

Figure 7 Comparative word accuracy for a choice of sub-band systems

Indeed the 5 sub-bands system appears to be the most efficient : in the following study we will therefore concentrate on a hybrid based on this more efficient 5 bands model.

The next step is training both a "clean" and "noisy" model to combine into one. The clean model is fed clean data according to the baseline system experimental building protocol, the noisy model is fed with the same set of training data, however a stationary white noise of SNR 10dB has been added to all utterances.

For this matter a tool had to be built/modified (this was to be of use when creating sets of artificial test data) that could "add" two waveforms according to a specified SNR.

After doing so, the clean 5 bands-model was benchmarked using the same experimental conditions as the baseline; the noisy model was benchmarked using a clean test data set corrupted once again with white noise of SNR 10dB. The results are shown in table 3 and 4. To have a means of comparison for the 5 bands noisy model, 2 3 and 7 band models were also created/benched following the same principle below.

Test set A					
	Subway	Babble	Car	Exhibition	Average
Clean	90.30	90.02	88.73	89.94	89.75
20 dB	78.35	87.18	85.62	81.39	83.14
15 dB	73.20	83.40	79.33	75.13	77.77
10 dB	64.54	73.40	65.11	61.86	66.23
5 dB	46.15	50.85	37.76	41.53	44.07
0 dB	24.16	23.79	15.90	22.59	21.61
-5dB	11.15	11.28	9.07	10.86	10.59
Average	55.41	59.99	54.59	54.76	

Test set B					
	Restaurant	Street	Airport	Station	Average
Clean	90.30	90.02	88.73	89.94	89.75
20 dB	85.32	84.13	86.70	84.05	85.05
15 dB	81.64	78.60	82.52	81.24	81.00
10 dB	73.50	67.50	72.53	71.77	71.83
5 dB	52.93	45.10	52.13	48.44	49.65
0 dB	28.34	23.10	27.26	20.86	24.89
-5dB	12.25	12.30	12.38	11.45	12.10
Average	60.61	57.25	60.82	58.25	

Table 3 : 5bands recognition results for clean training

2Bands	**92.38**
3Bands	**78.94**
5Bands	**80.41**
7Bands	**77.46**

Table 4 : recognition results for noisy training for a choice of sub-band models

Of course we observe here that the performance of the 5 bands system is a little lower than that of the baseline performance depicted in the baseline section of this report. Effort will now be made to improve the performance of the 5 bands system so that it outperforms the baseline. For this matter we will make use of the Bayesian Network framework further explained in the next section.

Hybrid HMM/BN Framework

In our hybrid HMM/BN framework, each acoustic model is a hidden Markov model (HMM) . The advantage is that on one hand all current HMM algorithms on speech recognition can still be used, and on the other hand the Bayesian networks allows us to incorporate an exterior, acoustic dependent information: here, whether the speech frame is considered noisy or not. We illustrate this below. From now on, circles denote continuous nodes, squares denote discrete nodes, clear means hidden, shaded means observed [5].

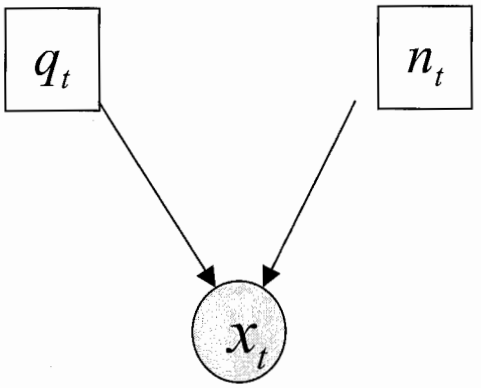

Figure 8 A Bayesian network relationship between observation x, state q, and noise factor n.

The noise factor, n_t, is a discrete quantity describing the noise conditions, for instance in a simple way $n_t = 0$ for clean or $n_t = 1$ for noisy .

A note on Dynamic Bayesian Networks (DBNs) also referred to as Temporal Bayesian Networks : DBNs are directed graphical models of stochastic processes. They generalize hidden Markov models (HMMs) by representing the hidden (and observed) state in terms of state variables, which can have complex interdependencies. The graphical structure provides an easy way to specify these conditional independencies, and hence to provide a compact parameterization of the model.

The simplest kind of DBN is a Hidden Markov Model (HMM), which has one discrete hidden node and one discrete or continuous observed node per slice. For instance, for a sequence of three states we would have :

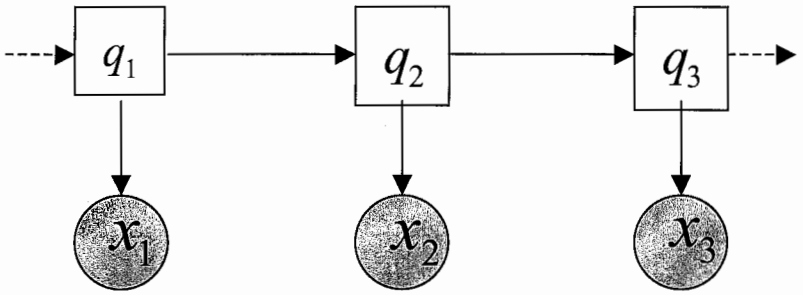

Figure 9 Classic 3-state HMM sequence of gaussian output.

In our hybrid HMM/BN framework, the general scheme will therefore be :

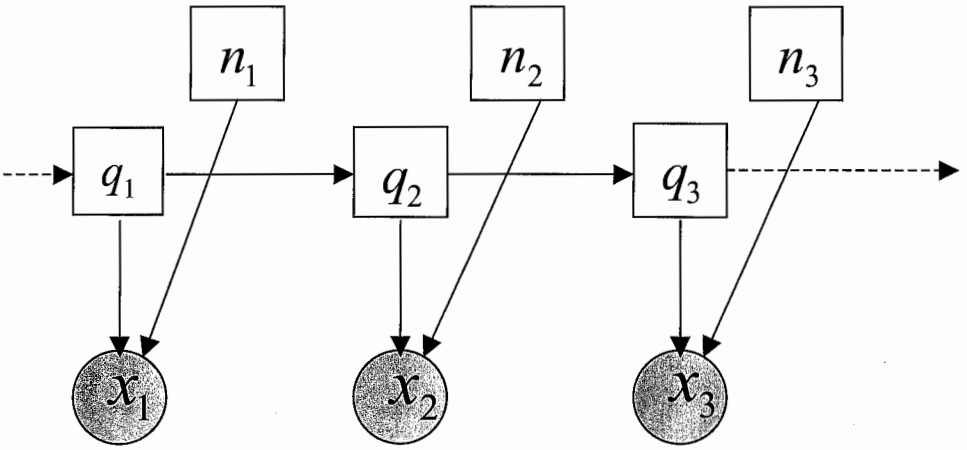

Figure 10 A 3-state hybrid HMM/BN

The presence of a noise factor n_t can be detected by computing the posterior probability of n_t for an observation x_t ; the probability of a frame being noisy or not is therefore:

$$P(n_t = k \mid x_t) = \frac{P(n_t = k, x_t)}{P(x_t)}$$

$$= \frac{\sum_{q_t \in Q} P(n_t = k, q_t, x_t)}{P(x_t)}$$

However, according to the dependencies in the Bayesian network described in Fig. X, the joint probability

$$P(n_t = k, q_t, x_t) = P(q_t)P(n_t = k \mid q_t)P(x_t \mid q_t, n_t = k)$$

can be reduced to

$$P(n_t = k, q_t, x_t) = P(q_t)P(n_t = k)P(x_t \mid q_t, n_t = k)$$

Therefore we can express :

$$P(n_t = k \mid x_t) = \frac{\sum_{q \in Q_t} P(x_t \mid n_t = k, q_t)P(q_t)P(n_t = k)}{P(x_t)}$$

At this point it will be interesting to evaluate $P(n_t = k / x_t)$ progressively, as we will develop further on, when we tamper with the way the recognizer deals with calculating the probability that such frame (observation) sub band be noisy or not.

Building the hybrid system

At first we will begin by assuming that the knowledge of $\sum_{q \in Q_t} P(x_t \mid n_t = k, q_t)$ is enough to let us decide if such frame (observation) sub band is noisy or not.

1st Comparison : $\sum_{q \in Q_t} P(x_t \mid n_t = 0, q_t)$ **and** $\sum_{q \in Q_t} P(x_t \mid n_t = 1, q_t)$

The channel considered noisy will therefore be marginalized for the selected frame (observation vector), that is, its computational weight will be set to 0 :

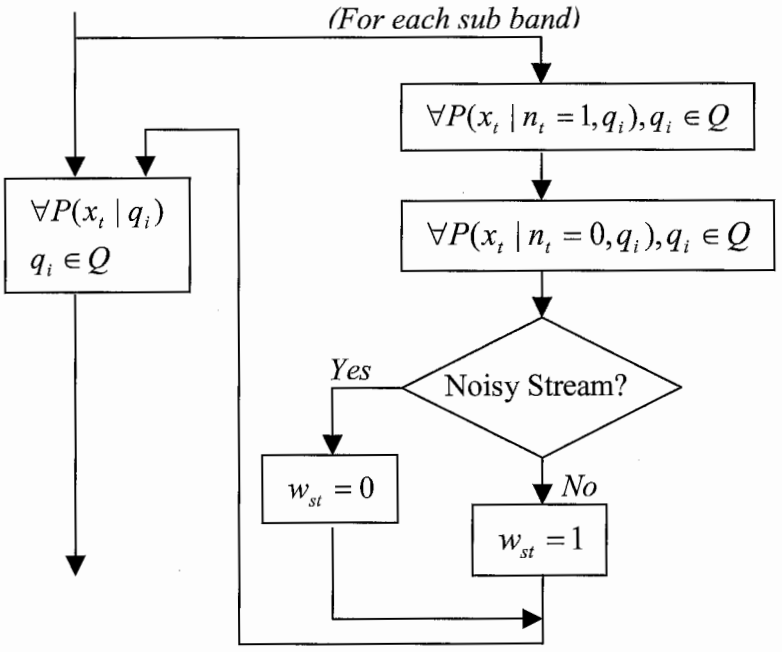

Figure 11 Assigning channel (sub band) weights

Then to improve our results and reduce the False Acceptance / False Rejection percentage we will include the information of $P(q_t)$ based on our knowledge of the previous frame (observation) :

$$P(q_t) = P(q_{t-1} \mid x_{t-1}) = \frac{P(x_{t-1} \mid q_{t-1})}{\sum_{q \in Q_t} P(x_{t-1} \mid q_{t-1})} \text{ (with } P(q_t) = \frac{1}{Q} \text{ for the first frame)}$$

Therefore we now have :

2nd Comparison : $\sum_{q \in Q_t} P(x_t \mid n_t = 0, q_t) \frac{P(x_{t-1} \mid q_{t-1})}{\sum_{q \in Q_t} P(x_{t-1} \mid q_{t-1})}$ **and**

$$\sum_{q \in Q_t} P(x_t \mid n_t = 1, q_t) \frac{P(x_{t-1} \mid q_{t-1})}{\sum_{q \in Q_t} P(x_{t-1} \mid q_{t-1})}$$

Lastly we will include the information of $P(n_t = k)$:

$$P(n_t = k) = \frac{P(n_t = k \mid x_{t-1})}{\sum_{k \in N} P(n_t = k \mid x_{t-1})} \quad \text{(with } P(n_t = k) = 0.5 \text{ for the first frame)}$$

Therefore we now have :

3rd Comparison : $\sum_{q \in Q_t} P(x_t \mid n_t = 0, q_t) \frac{P(x_{t-1} \mid q_{t-1})}{\sum_{q \in Q_t} P(x_{t-1} \mid q_{t-1})} \frac{P(n_t = 0 \mid x_{t-1})}{\sum_{k \in N} P(n_t = k \mid x_{t-1})}$ **and**

$$\sum_{q \in Q_t} P(x_t \mid n_t = 1, q_t) \frac{P(x_{t-1} \mid q_{t-1})}{\sum_{q \in Q_t} P(x_{t-1} \mid q_{t-1})} \frac{P(n_t = 1 \mid x_{t-1})}{\sum_{k \in N} P(n_t = k \mid x_{t-1})}$$

Thresholds have to be set in order to equalize the False Acceptance / False Rejection percentages : the following graphical method is used, data consisting of 5 male voice utterances and 5 female voice utterances in clean speech for the false rejection rate, and the same data corrupted with white noise SNR 10dB for the false acceptance rate.

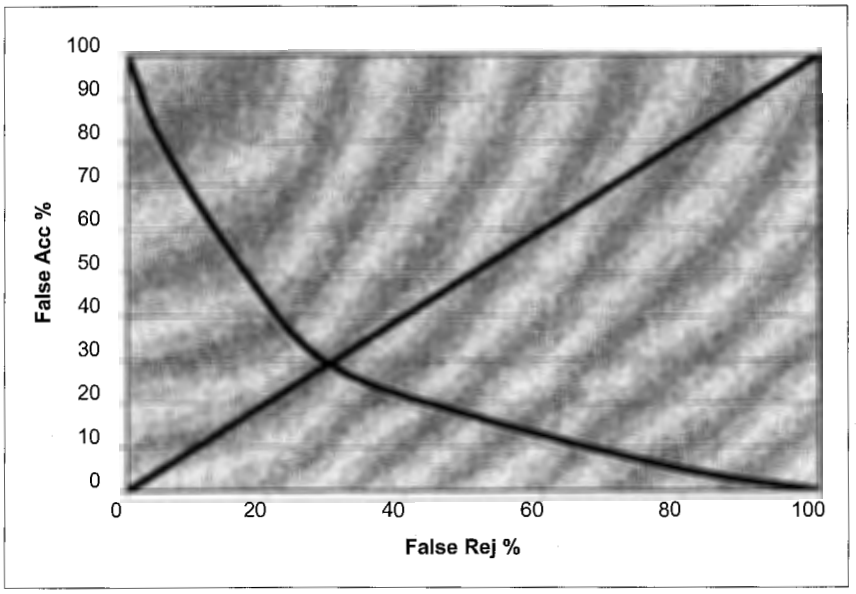

Figure 12 Threshold setting method

The improvements from stage to stage are listed in the following table :

Sub-band	1st Stage	2nd Stage	3rd Stage
1	30	30	21
2	35	37	29
3	38	38	38
4	34	22	30
5	20	22	8

Table 5: False Acc./False Rej. rates (in %) through the different system evolutions

Tests on artificial data

As a means of testing we now find it useful to construct some artificial testing data : we will use as described in [1] some clean speech data along with additive white-noise passed through a band-pass filter with a 3dB cutoff-bandwidth (stationary band-selective noise) and a varying central frequency. We will therefore use 4 sets of relevant data :

- Clean data
- 1 band corrupted (central frequency 600 Hz)
- 2 bands co corrupted (central frequency 600 &1200 Hz)
- 3 bands corrupted (central frequency 600, 1200 & 2000 Hz)

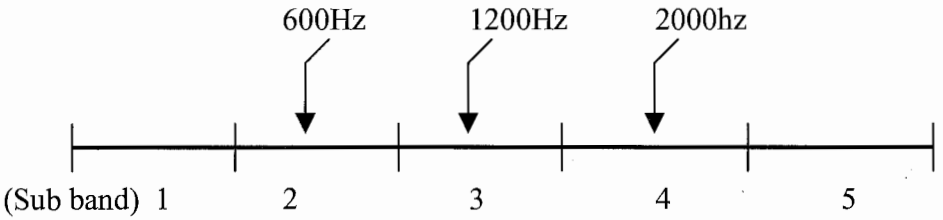

Figure 13 artificial data spectral corruption

Furthermore, I implemented an algorithm to reject only the 3 bands considered most noisy : this is in case that 4 bands or more over the 5 are considered noisy, and to therefore avoid the total lack of information to reach the recognizer.

Word Acc. (%)	Full band	5 Sub bands (oracle)	1st stage Hyb.	2nd stage Hyb.	3rd stage Hyb
Clean	97.97	92.88	92.31	89.75	90.2
1 Band corruption	18.79	92.31	23.93	27	34.3
2 Bands corruption	12.77	87.18	5.7	3.5	4.42
3 Bands corruption	10.19	70.94	9.69	7.98	8.75

Table 6: Test on artificial data

It appears here that the good results from the 5 bands "oracle" (i.e. we know which band is corrupted or not prior to the weighting of the bands) are far from being reached when band-selection is supposed to be automatic, although a significant improvement is noticeable in the case of 1 band corruption over the full band baseline. Further experiments have also shown that the more corrupted the band is (i.e. with a more than 100Hz 3dB cutoff frequency) the more accurate the detection, and therefore the more improvement we get (from 50% to 80% Word Accuracy for single band corruption, depending on which band is corrupted).

Tests on real data

Tests were conducted as previously mentioned on AURORA 2.0 standard test sets a and b. As a means of comparison we have included previous results from the 2 and 5 sub-band systems.

The results are shown in the figure below. Detailed result charts are included in Appendix II.

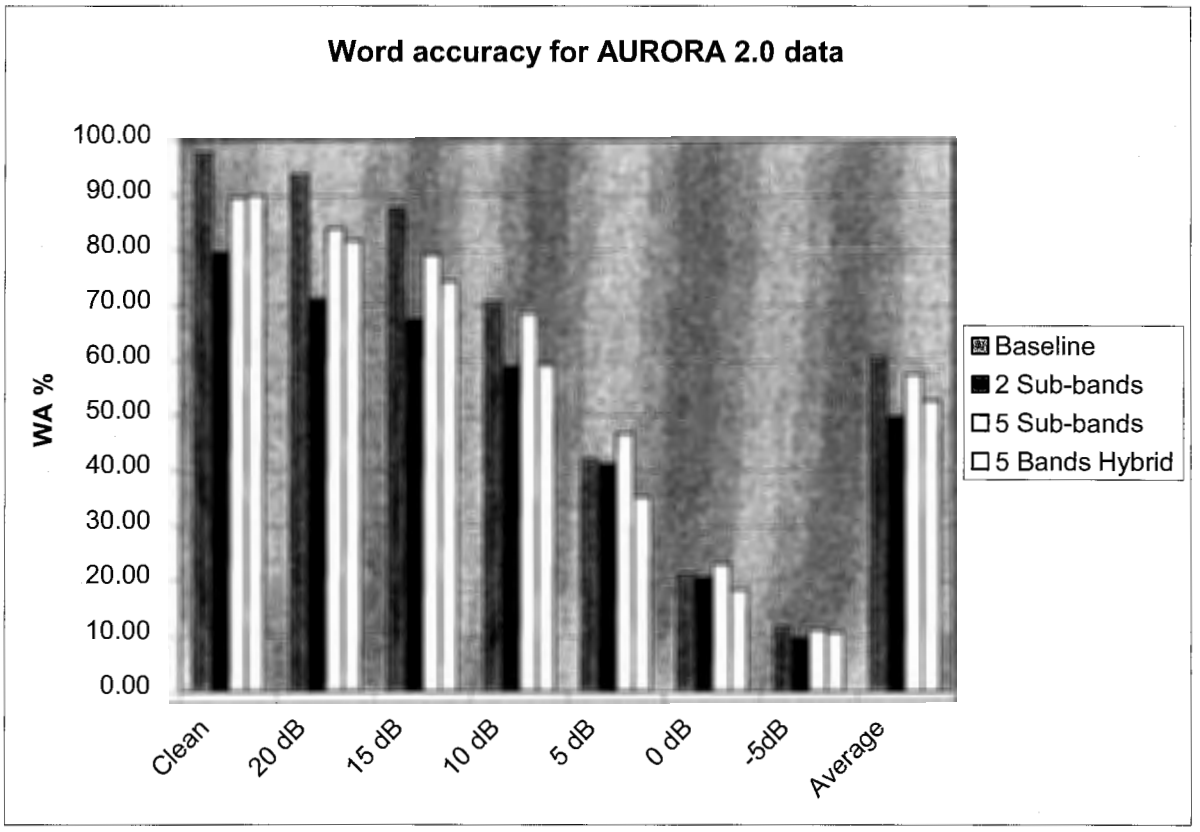

Figure 14Comparative Word Accuracy for "real" data testing

As we can observe, the 5 band hybrid never surpasses the baseline system : it appears that band selection accuracy is not high enough to allow accurate selection of which channel contains valuable information, and which one does not : False Acceptance / False Rejection rates are still too high to operate properly on real data, even though to some extent results were acceptable in the case of "ideal" artificial corruption.

Further Improvements

Several improvements over the current methods should be considered for future work, both in the fields of feature extraction and noise detection.

Concerning feature (MFCC) extraction, it has been suggested to make use of sub-band overlapping. The composite spectrum of all filters (filter bank of triangular filters) should be more or less flat. However, instead of assigning such series of filters to a band and then the following filters to the next band without overlapping (what we have done here) , allow some filters at the border of each contiguous band to be part of both bands. For instance, instead of having a 30 filter bank , 5 bands system assigned as 6/6/6/6/6, we could allow more than 6 filters to each band given the composite spectrum remains flat. In any case, cepstral coefficients always do the overlapping/smoothing of the spectrum

Concerning noise detection, in my opinion extensive study has to be conducted on which noisy trained model to use in order to achieve more accurate results, that is on the nature of the noise used to train the noisy model. While using stationary SNR10dB white noise seemed the most obvious option, tests have shown that detection rates tend to be go down even though same SNR evenly distributed noise is still the source of corruption, although on a tighter band. The "tolerance" of noise type variation definitely needs improvement over the present state to lower false acceptance/rejection rates (around 30% is just too much).

Global Conclusion

As a conclusion, I would say that the work conducted here was quite useful not in the fact that it shows any improvement over the present method overall (because in fact, it does not), but because it pin points more precisely what efforts and improvements have to made in the field of noise detection to achieve greater robustness of the overall system. The hybrid HMM/Bayesian network method/framework is correct and shows great promise given these problems can be even partly solved ; this is comforted by the fact that on ideal noise detection (band marginalization), results are very good. Therefore, one should not consider that lack of WA improvement in automatic speech recognition research is "waste of time". This challenging research topic opens the door to a lot of future investigations on noise detection.

On a personal point of view, this internship period at ATR was my first experience in the field of research. This allowed me to discover a whole new universe that was unknown to me before, to practice my academic skills as well as my few previous experiences practical engineering skills, but also to witness the fact that research is a great federative of human collaborations, cultures, languages from around the world. Moreover, this has given me the will to continue deeper studies in the field of Automatic Speech Translation/Recognition in the near future.

Working in Japan was overall a great experience both professionally and personally : it is fascinating to see how one of the most advanced country in state of the art technologies manages to combine its own research power and knowledge with other countries own to advance even further. I look forward to coming back soon to work in Japan, and in ATR .

References

Publications

[1] Ming (J.), Jancovic (P.) and Smith (F.J.)
Robust speech recognition using probabilistic union models
Proc. Of IEEE transactions on speech and audio processing, VOL. 10, NO. 6, SEPTEMBER 2002

[2] Ming (J.) and Smith (F.J.)

A posterior union model for improved robust speech recognition in non-stationary noise
Proc. Of ICASSP 2003

[3] Daoudi (K.), Fohr (D.) and Antoine (C.)
Continuous multi-band speech recognition using Bayesian networks
Proc. Of IEEE ASRU. Trento, Italy, December 9 -13th, 2001

[4] Hermansky, (H.), Tibrewala, (S.) and Pavel, (M.),
Towards ASR on partially corrupted speech
PROC .Of ICSLP96, October 1996

Technical Report

[5] Mak (B.), Markov (K.) and Nakamura (S.)
Clean and noisy speech detection using hybrid HMM/BN framework
SLT-TR, internal use

Miscellaneous articles / books

[6] Hirsch (H-G.) and Pearce (D.)
The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions

[7]Rabiner (L.) and Juang (BH.)
Fundamentals of Speech Recognition
Englewood Cliffs, NJ: Prentice-Hall International Editions, 1993

[8] Young (S.) and al.
The HTK Book
Revised for HTK version 2.2, January 1999

[9] Murphy (K.)
A brief introduction to graphical models and Bayesian networks, 1998

Appendix 1 : Mel-filterbank analysis

The human ear resolves frequencies non-linearly across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves recognition performance. A popular alternative to linear prediction based analysis is therefore filterbank analysis since this provides a much more straightforward route to obtaining the desired non-linear frequency resolution.

However, the problem is to space the filters along the critical band in order to choose bands that give equal contribution to speech articulation.

The Mel-scale is a variant of the critical band scale and is defined by :

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

In order to implement this filterbank, the window of speech data is transformed using a Fourier transform and the magnitude is taken. The magnitude coefficients are then binned by correlating them with each triangular filter. Here binning means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results accumulated.

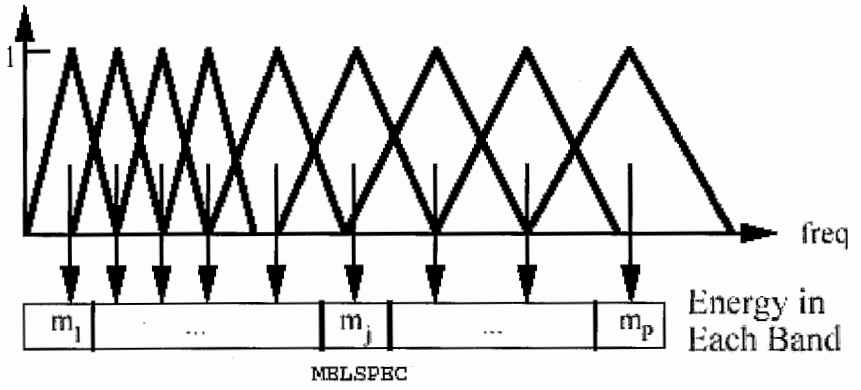

Figure 15 Mel filterbank analysis

Appendix 2 : AURORA 2.0 tests results

- *Baseline WA, full band 20 cepstral coefficients*

	Set a				Set b				
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Average
Clean	97.97	97.88	97.79	97.78	97.97	97.88	97.79	97.78	97.86
20 dB	95.24	93.11	95.68	95.65	91.53	95.13	91.14	93.18	93.83
15 dB	89.50	88.12	88.58	88.40	86.71	87.27	85.24	88.03	87.73
10 dB	72.21	75.12	67.91	63.25	74.12	66.51	73.99	73.16	70.78
5 dB	40.84	50.39	33.55	30.64	49.22	39.12	48.08	44.89	42.09
0 dB	21.77	23.13	18.70	18.08	23.15	22.37	23.41	20.24	21.36
-5dB	14.09	10.94	12.08	11.76	11.24	13.85	11.06	11.05	12.01
Average	61.66	62.67	59.18	57.94	61.99	60.80	61.53	61.19	

- *2 bands system, 28 cepstral coefficients*

	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Average
Clean	81.61	79.53	78.50	79.17	81.61	79.53	78.50	79.17	79.70
20 dB	72.58	69.89	78.65	58.38	66.44	74.00	75.25	74.95	71.27
15 dB	70.19	63.75	74.47	56.06	62.82	69.71	72.11	71.27	67.55
10 dB	63.52	58.37	61.23	47.21	57.02	57.89	64.33	62.48	59.01
5 dB	47.37	41.69	38.59	32.21	40.34	38.72	46.20	43.97	41.14
0 dB	26.04	21.01	16.05	15.89	22.78	19.20	25.53	18.30	20.60
-5dB	11.94	9.37	7.87	9.04	9.79	9.61	11.33	9.41	9.80
Average	53.32	49.09	50.77	42.57	48.69	49.81	53.32	51.36	

- *3 bands system, 30 cepstral coefficients*

	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Average
Clean	78.54	78.63	77.60	79.30	78.54	78.63	77.60	79.30	78.52
20 dB	36.11	53.05	60.57	38.60	48.36	56.74	61.17	61.09	51.96
15 dB	33.96	43.17	48.34	32.71	39.09	49.06	52.91	52.79	44.00
10 dB	29.32	32.44	34.18	28.32	29.60	40.72	39.40	38.72	34.09
5 dB	19.96	18.59	18.94	19.32	16.92	26.93	23.89	21.54	20.76
0 dB	12.47	10.79	10.92	11.67	10.22	13.94	13.27	10.03	11.66
-5dB	8.54	8.10	8.20	8.98	7.65	9.04	8.86	7.96	8.42
Average	31.27	34.97	36.96	31.27	32.91	39.29	39.59	38.78	

- *5 bands system, 30 cepstral coefficients*

	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Average
Clean	90.30	90.02	88.73	89.94	90.30	90.02	88.73	89.94	[illegible]
20 dB	78.35	87.18	85.62	81.39	85.32	84.13	86.70	84.05	[illegible]
15 dB	73.20	83.40	79.33	75.13	81.64	78.60	82.52	81.24	[illegible]
10 dB	64.54	73.40	65.11	61.86	73.50	67.50	72.53	71.77	[illegible]
5 dB	46.15	50.85	37.76	41.53	52.93	45.10	52.13	48.44	[illegible]
0 dB	24.16	23.79	15.90	22.59	28.34	23.10	27.26	20.86	[illegible]
-5dB	11.15	11.28	9.07	10.86	12.25	12.30	12.38	11.45	[illegible]
Average	55.41	59.99	54.50	54.76	60.61	57.25	60.32	58.25	

- *5 bands **Hybrid system**, 30 cepstral coefficients*

	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Average
Clean	90.14	90.19	89.23	89.82	90.21	91.88	89.32	89.82	[illegible]
20 dB	74.85	85.80	83.30	78.90	84.19	81.68	84.37	82.10	[illegible]
15 dB	69.36	79.12	73.13	69.61	77.34	72.71	79.66	74.95	[illegible]
10 dB	61.16	63.78	52.85	54.00	63.17	57.25	64.24	59.06	[illegible]
5 dB	43.23	34.84	26.19	33.17	41.17	34.27	38.11	32.43	[illegible]
0 dB	23.73	16.19	14.61	17.74	21.58	16.67	20.64	17.09	[illegible]
-5dB	12.77	9.87	10.41	9.26	12.65	10.38	11.72	9.56	[illegible]
Average	58.61	54.26	49.96	50.36	55.76	52.12	55.44	52.14	

- *7 bands system, 28 cepstral coefficients*

	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Average
Clean	88.15	87.91	86.85	87.50	88.15	87.91	86.85	87.50	[illegible]
20 dB	41.51	50.73	48.79	40.60	47.80	47.46	54.10	45.73	[illegible]
15 dB	36.48	40.05	41.04	31.13	35.25	38.66	42.41	36.35	[illegible]
10 dB	27.33	22.40	27.05	17.06	20.60	24.12	27.74	23.63	[illegible]
5 dB	15.29	8.62	8.71	4.54	5.07	9.49	12.02	7.19	[illegible]
0 dB	9.67	3.39	3.10	2.75	2.70	4.53	0.42	1.02	[illegible]
-5dB	5.80	3.36	4.44	3.09	1.38	3.48	0.30	1.76	[illegible]
Average	32.03	30.92	31.43	26.67	28.71	30.81	31.98	29.03	