NOISE ROBUST ESTIMATE OF SPEECH DYNAMICS FOR SPEAKER RECOGNITION

J.P. Openshaw[1] *& J.S. Mason*

Speech Research Group, Dept of Electrical Engineering, University College of Swansea,
Swansea, SA2 8PP, UK Tel: +44 1792 295422 Fax: +44 1792 295686
1. Currently with the SVR group, Cambridge University Engineering Dept. Trumpington Road,
Cambridge, UK.

ABSTRACT

This paper investigates the robustness of cepstral based features with respect to additive noise, and details two methods of increasing the robustness with minimal need for *a-priori* knowledge of the noise statistics.

The first approach is a form of noise masking which adds a fixed offset to the linear spectral estimate.

The second is a form of sub-band filtering, again in the linear domain, which estimates the dynamic content of the speech using Fourier transforms. This avoids negative values normally inherent in such filtering and which presents difficulties in deriving log estimates.

Both methods are shown to provide useful levels of robustness to additive noise, for example, speaker identification error rates in SNR mis-matched conditions of 15 dB are reduced from 60.5% for standard mel cepstra to 13.8% and 24.1% for the two approaches respectively.

1 INTRODUCTION

Static features, computed from a single windowed frame of speech, attempt to capture the instantaneous (assumed quasi-stationary) spectral characteristics along the time course. Dynamic features (velocity and acceleration) attempt to capture the dynamics of the speech by processing a short sequence of static features. A similar idea of relative spectra is the motivation behind Hermansky's RASTA [3], which also aims to capture changes in spectra. The associated processing involves sub-channel filtering, and when applied in the *linear* spectral domain gives rise to practical difficulties in computing the cepstra. Following the filtering, negative values can and normally will occur inhibiting the conventional log function evaluation. Hermansky overcomes this problem by adding a fixed offset to the output and notes an increase in robustness to additive noise. This raises the interesting question of the level of offset and Hermansky shows that the optimum lies between the level of noise and the level of the speech signal. We demonstrate that the mere inclusion of the offset itself has a major beneficial effect in reducing sensitivity of cepstra to additive noise.

We also propose a simple magnitude Fourier approach which provides an alternative solution to the negative spectra problem and at the same time gives dynamic parameters from a short sequence of static frames. The features are termed Fourier sub-band filtered coefficients (F-SBF).

1.1 Experimental Base

As a benchmark, the effect that noise has on the recognition performance is shown in Figure 1. Here the task is speaker identification (SI) and the experimental conditions are as follows.

The population is 20 speakers taken from the BT CONNEX database. An alphabet vocabulary is modelled using a text-independent VQ classifier consisting of 32 centroids, and a weighted Euclidean distance metric is used throughout. Features are MFCC-14 from a frame size of 25.6 ms, with a 50% overlap. The linear regression Δ features are as specified in [2], and computed over 6 static frames. Gaussian noise is added to the clean train/test data.

Three experiments are performed, 'cn', 'nc' and 'nn'. The first letter indicates whether clean or noisy data is used in training, with the second letter indicating the conditions used in testing. When noise is added to either the training or test data, the level is as indicated on the abscissa. The 'cn' and 'nc' experiments relate to cross-testing, while the 'nn' profiles relate to matched noise conditions.

It is seen that for the cross-test experiment, cn, errors rise rapidly even in moderate amounts of noise, yielding errors of *circa* 60% at an SNR of only 15dB. Comparing this to the performance in clean conditions, with an error rate of just 3.4%, it can be seen that the effects of noise are marked.

In comparison, the 'nn' case where the models are trained using *a-priori* knowledge of the noise conditions in testing shows a greater degree of robustness, for example when the test speech has an SNR of 15dB, the recognition error rate is just 12.9%.

The Δ-MFCCs, give similar results to those of the static feature but with the Δ features generally faring worse in high SNR conditions.

2 STATIC NOISE MASKING

2.1 Overview

The technique of noise masking, originally investigated by Klatt [6], can be viewed as the antithesis of spectral subtraction techniques. The basic tenet of operation is to maintain parity between training and test phases, with noise accepted as an inevitable circumstance. Klatt artificially maintains a minimum power in the filter bank, representative of the noise spectrum found in training. More recently Varga [9] provides
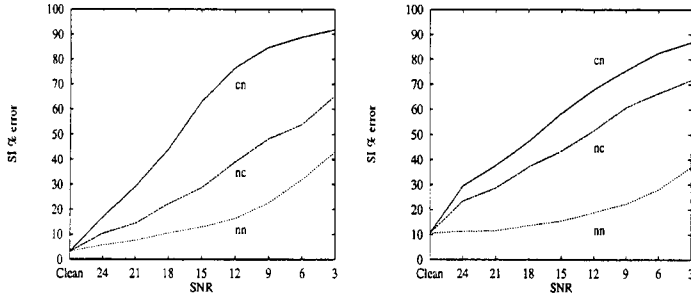
Figure 1: MFCC(left) and Δ-MFCC(right) SI recognition error rates. cn: clean model, noisy test data, nc: noisy model, clean test data, nn: noisy model and noisy test data. The SNR is as indicated on the abscissa.
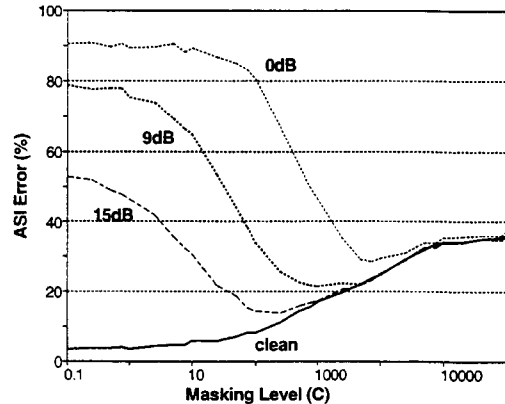

Figure 2: SI recognition error rate for the static noise masking technique. The masking level (C) is as indicated on the abscissa, clean training models are used in all cases, and the test SNR is as indicated on each individual profile.

a study of various noise masking algorithms, with Mellor [7] providing equivalent effects but masking in a transform domain, as opposed to a log spectral one.

Here we examine the limits of noise masking by adding a fixed constant to each band-pass (mel) filter, prior to the log function in the cepstral analysis. Thus the cepstra is evaluated from

$$cepstra = \mathcal{F}_{DCT}(log(C + S(w)))\qquad(1)$$

where $\mathcal{F}_{DCT}$ represents the cosine transform, $S(w)$ the speech spectra, and C the masking level. The motivation is to lessen the sensitivity of cepstra to additive noise by a swamping effect.

A direct comparison can be drawn with Hermansky's extension to RASTA processing of the PLP feature [4]. To overcome the difficulty of applying log-like non-linearities to the band-pass filtered power spectrum (which could have negative values following the filtering), Hermansky proposes the approximation

$$cepstra = \mathcal{F}_{DCT}(log(1 + J.S(w)))\qquad(2)$$

He then shows that the constant J can be designed to minimise noise sensitivity, although the contribution of J and of the RASTA processing itself is unclear. It is obvious that Hermansky's J constant and our noise masking (fixed) offset are directly equivalent.

2.2 Experimental Results

Figure 2 displays the SI error rate when testing under clean conditions, and that of 15, 9 and 0db SNR versus masking level. In all cases, clean models are used.

The most significant result of Figure 2 is the improvement with the addition of masking, even with a test SNR as low as 0db. As can be expected, a larger masking level C is needed to combat a lower SNR. It can also be seen that the test SNR is fairly insensitive to the masking level used, ie a single masking level caters for a range of test SNRs. The optimum masking level is SNR dependent, and a conflict of two contrary effects.

The approach alters the log non-linearity and in this sense is similar to the root homomorphic work of Alexandre [1] and the work of Hermansky. Unfortunately it is shown empirically that altering the non-linearity is a double-edged sword in terms of reducing the sensitivity to changes in noise, as

it also reduces the sensitivity to the content of the speech, increasing the errors. It also removes the level independence provided by the log function with a single term.

The effect of altering the non-linearity is displayed in a simple manner in Figure 3. The top-right graph has no masking and shows the output of 32 mel-scaled filters for the utterance "a" (/ey/). The following 3 graphs each include an increasing masking level, from C=100, which is an approximate optimum for recognition at 15dB, and for masking levels of 1000 and 10000 respectively. The three profiles on each graph represent clean, 6 and 0dB SNR from the bottom up.

As the masking level is increased, the profiles for different SNRs tend to converge, reducing the spectral distance between mis-matched SNR conditions. Also there is a tendency for a heavy masking level to remove detail from the spectra. For instance, comparing the clean profile with no masking and with a masking level of 10000, a significant amount of the speech spectra is swamped by the masking spectra reducing the 'individuality' of the spectra. It is this effect that reduces the recognition accuracy in high masking levels.

3 ESTIMATING SPEAKER DYNAMICS USING SUB-BAND FILTERING

Although other techniques for measuring speech dynamics exist, such as Δ-cepstra [2], and the 2 dimensional cepstra based work of Kitamura [5] and Vaseghi [10], these tend to be affected in a complex manner by additive noise as they essentially operate on log spectral estimates [8].

This paper introduces a new form of sub-band filtering (SBF), based on Fourier analysis of sub-bands in the *linear* domain, and thus termed Fourier sub-band filtering (F-SBF). The form of F-SBF used here estimates the dynamic content of the speech, assuming the noise to be stationary relative to the speech dynamics.

The important merits of the F-SBF approach are:
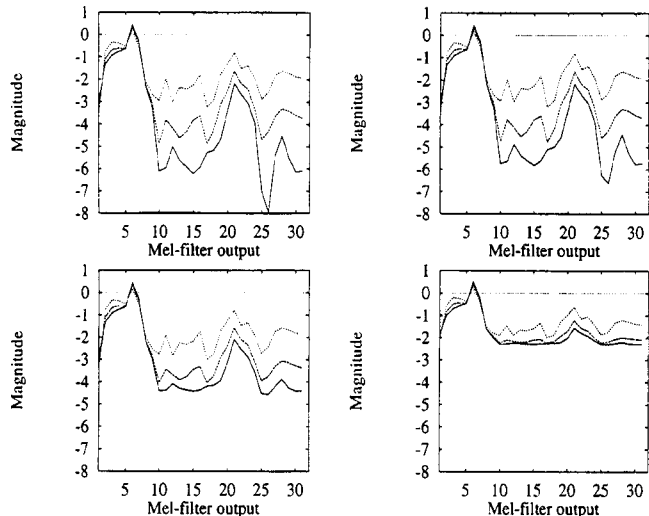
- no *a-priori* knowledge of the noise is needed,

Figure 3: The log mel-filter output for the a voiced part of the utterance "a". The top-left graph has no noise-masking, the following graph has a masking level optimum for approximately 15dB, $C = 100$, with the 2 following graphs increasing the masking level by an order of magnitude successively. The three profiles on each of the graphs are, in descending order, for a clean portion of the speech, and for SNRs of 6 and 0dBs respectively.

- additional computation (over standard cepstra) is minimal,

- other benefits of standard cepstra are retained: good overall performance, pitch filtering and a compact representation.

A practical difficulty of processing spectral estimates in the linear domain, when such estimates are to be converted to cepstra, is the requirement to retain positive values for the estimates, essential for the subsequent log function. This problem exists both with the popular spectral subtraction approach and the sub-band filtering such as J-RASTA, for example. Solutions include clamping and using a fixed offset.

Here, the Fourier output (magnitude term) overcomes this difficulty in providing estimates of dynamics which are inherently positive and can therefore be followed directly by the log and cosine operations.

Equation 3 defines the process:

$$c_j = \mathcal{F}_{DCT}.\log.\mathcal{F}_{mag}|S_i(w)| \qquad (3)$$

where c_j is the j'th feature coefficient, $S_i(w)$ is the set of filter outputs at time i, $\mathcal{F}_{mag}$ is the magnitude of the Fourier transform, $\mathcal{F}_{DCT}$ is the discrete cosine transform.

3.1   Experimental results

The same experimental conditions as detailed above apply. Figure 4 shows the recognition % error profiles for the F-SBF features, and also for standard MFCCs. In all cases, the models are trained in clean conditions, and tested with speech with an SNR as shown on the abscissa. Gaussian white noise is used.
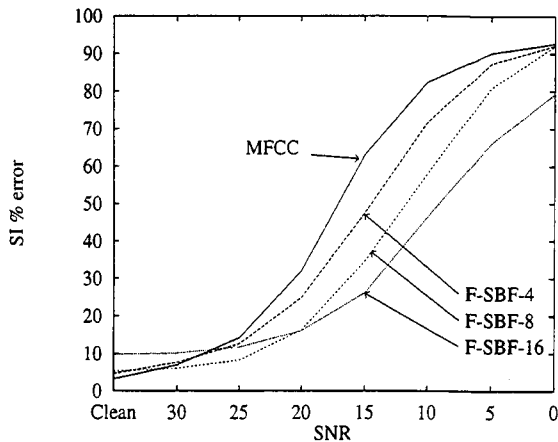

Figure 4: SI % error rates for MFCC and F-SBF. Clean models are used in all cases, with the test SNR as indicated on the abscissa. The 3 recognition profiles for F-SBF relate to the number of frames used to obtain the relative spectral estimate, in this case 4, 8 and 16 frames.

The three F-SBF profiles relate to the number of consecutive spectral observations used to obtain the estimate of the dynamic content in this case 4, 8 and 16.

The F-SBF features clearly out-perform standard MFCCs in cross-test conditions. For example, at 15dB SNR the cross-test MFCC error rate is 60.5%, the F-SBF error rate is 24.1%, using 16 spectral observations to measure the dynamic content. As is noted with the static noise masking technique, a compromise between performance in clean and noisy conditions has to be made. Longer windows result in a feature less responsive to noise, but which do not perform as well in clean conditions. Interestingly, using dynamic features with short windows results in a performance in clean conditions similar to that obtained by using standard MFCC features.

3.2   Pitch response of the F-SBF feature

One of the useful characteristics of standard cepstral analysis is attenuation of the pitch component.

Figure 5 shows the effect of pitch on both the mel-based F-SBF feature and standard MFCCs. The full feature order is used in both cases. For the F-SBF feature, the parameters used are 256 sample window, with 128 sample overlap and sub-band filtering is performed over 8 frames. The coefficient is denoted on the x-axis, with the pitch impulse frequency on the y-axis. The value of the coefficient is indicated by the z-axis.

The comparison between Figure 5 left (F-SBF) and right (MFCC) shows that the F-SBF feature is comparable, in terms of sensitivity, to standard MFCCs upto a pitch frequency of approximately 200 Hz. Pitch frequencies exceeding 200Hz have a major effect on the higher order coefficients of standard mel-based analysis. These effects are attenuated in the mel-based F-SBF feature, permitting the use of a wide range of coefficients without pitch influence.
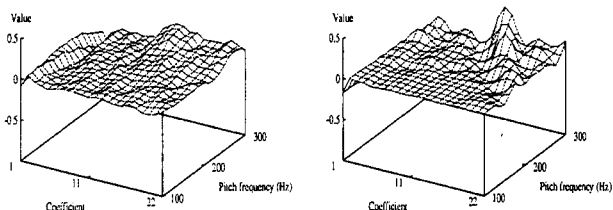
Figure 5: The effects of pitch on the mel-based F-SBF feature (left) and standard MFCCs (right). The pitch signal is represented by single impulses chirped between 100–300 Hz. The F-SBF parameters are over 8 frames.

4 CONCLUSIONS

This paper has demonstrated the adverse effects of additive noise on cepstral based features by examining recognition performance. A technique called static noise masking attempts to make the features more inherently robust, with minimal need for a-priori knowledge of the noise statistics.

The conclusions of static noise masking is that altering the non-linearity is beneficial in terms of increasing the robustness to additive noise, with the cross-test errors reduced from 60.5% to 13.8% for an SNR of 15dB.

Although masking levels are to some extent SNR dependent a low sensitivity to a sub-optimal masking value is demonstrated.

A form of Fourier sub-band filtering in the linear spectral domain is demonstrated. This feature is unaffected by additive spectral components constant over the window duration, but possibly variable otherwise, for example between test and training.

A trade-off between robustness to noise, and performance in clean conditions can be made by altering the number of frames over which the relative estimate is made: estimating using more frames is more robust to noise, while using fewer frames gives better performance in clean conditions. A directly comparable cross-test SI error rate is reduced from 60.5% for MFCCs to 24.1%.

REFERENCES

[1] P. Alexandre, J. Boudy, and P. Lockwood. Root homomorphic deconvolution schemes for speech processing in car noise environments. Proc. ICASSP-93, pages 99–102, 1992.

[2] S. Furui and A. E. Rosenberg. Experimental studies in a new automatic speaker verification system using telephone speech. Proc. ICASSP-80, pages 1060–1062, 1980.

[3] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). Proc. Eurospeech-91, pages 1367–1370, 1991.

[4] H. Hermansky, N. Morgan, and H. Hirsch. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. Proc. ICASSP-93 Vol. II, pages 83–86, 1993.

[5] T. Kitamura and S. Imai. Pitch determination by two-dimensional cepstrum. Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology, 37:25–32, March 1976.

[6] D. H. Klatt. A Digital Filter Bank For Spectral Matching. Proc. ICASSP-76, pages 573–576, 1976.

[7] B. A. Mellor and A. P. Varga. Noise masking in a transform domain. Proc. ICASSP-93, pages 87–90 Vol. II, 1993.

[8] J. P. Openshaw. The effects of additive noise in speaker recognition. Ph.D. Thesis, University College Swansea, 1995.

[9] A. P. Varga and K. Ponting. Control experiments on noise compensation in hidden markov model based continuous word recognition. Proc. Eurospeech-89, pages 167–170, September 1989.

[10] S. V. Vaseghi, B. P. Milner, and P. N. Conner. Speech modelling using cepstral-time feature matrices in hidden Markov models. Proc. IEE-I, 140, No. 5:317–320, October 1993.