AUDITORY MASKING BASED ACOUSTIC FRONT-END FOR ROBUST SPEECH RECOGNITION

K.K. Paliwal and B.T. Lilly

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
B.Lilly, K.Paliwal@me.gu.edu.au

ABSTRACT

This paper presents an acoustic front-end which uses the properties of auditory masking for extracting acoustic features from the speech signal. Using the properties of simultaneous masking found in the human auditory system, we compute a masking threshold as a function of frequency for a given speech frame from its power spectrum. All those portions of the power spectrum which are below the auditory threshold are not heard by the human auditory system due to masking effects and hence can be discarded. These portions are replaced by the corresponding portions in the masking threshold spectrum. This modified power spectrum is processed by the linear prediction analysis or homomorphic analysis procedure to derive cepstral features for each speech frame. We study the performance of this front-end for speech recognition under noisy environments. This front-end performs significantly better than the conventional linear prediction or homomorphic analysis based front-ends for noisy speech. In terms of signal-to-noise ratio, simultaneous masking offers an advantage of more than 5 dB over the LPCC front-end in isolated word recognition experiments and 3dB in continuous speech recognition experiments.

1. INTRODUCTION

Most of the speech recognizers reported in the literature use cepstral features which are derived from the speech signal by using the linear prediction (LP) analysis technique [1]. These cepstral features are known to be very sensitive to additive noise and channel mismatch distortions which are very common in practical environmental conditions. As a result, the performance of these recognition systems deteriorates drastically in the presence of these types of distortions. On the contrary, human listeners can recognize speech even in the presence of large amounts of noise and channel distortions. Therefore, it is argued that the acoustic front-end can be made more robust to these distortions by utilizing the properties of human auditory system. We call these types of front-ends, auditory front-ends.

A number of auditory front-ends have been proposed in the literature. These front-ends utilise some property of the human auditory system, to modify the power spectrum and then use either the LP analysis technique or the homomorphic analysis technique to obtain the smooth spectral envelope, which is then represented in terms of a few cepstral features. Some examples of popular auditory front-ends are Mel filter-bank analysis [2], perceptual linear prediction (PLP) analysis [3] and ensemble interval histogram (EIH) analysis [4].

Mel filter-bank analysis warps the frequency scale such that it resembles the position along the basilar membrane in the human ear. Equivalently, the Bark frequency scale is calculated based on experiments that measure the frequency response of the basilar membrane in the human ear. The PLP front-end utilies several human auditory properties including a bark warped frequency scale. The EIH front-end models the auditory nerve system found in humans. These auditory front-ends have been studied in the past for speech recognition and their performance has been found to be better than the non-auditory front-ends, especially for noisy environments [4, 5].

Humans recognise speech containing large amounts of noise using two different types of masking: simultaneous and non-simultaneous masking. The properties of forward and backward non-simultaneous masking has been investigated and has been shown to show significant improvement for robust speech recognition[6, 7]. Here, we look at using the auditory property of simultaneous masking and apply it to speech recognition at various noise levels.

The following section gives a detailed description on how we use simultaneous masking to compute robust cepstral features for speech recognition. Section three describes the databases and the recognition setup used for the recognition experiments presented in this paper. Section four shows the results obtained when subjecting the masking front-end to white noise at various signal-to-noise ratio (SNR) levels on an isolated word and a continuous speech HMM recogniser. We also evaluate the performance of this front-end on a channel distortion filter at different levels. We compare this with using no masking to evaluate any increase in performance. A second set of experiments uses RASTA [6] as well as our simultaneous masking procedure to evaluate whether combining the two masking properties increases the performance of an isolated word HMM recogniser. Finally, section 5 gives some conclusions.

2. AUDITORY MASKING BASED ACOUSTIC FRONT-END

Humans use both simultaneous and non-simultaneous masking to recognise speech in adverse conditions such as

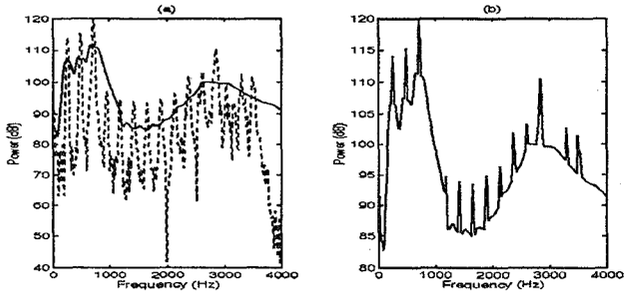

Figure 1. (a) Log power spectrum and masker. (b) New log power spectrum.

noise and channel distortions. [1] Non-simultaneous masking has two forms, forward and backward masking. Backward masking occurs when the signal precedes the masker in time. Forward masking occurs when the signal follows the masker and is utilised in techniques such as RASTA [6]. The RASTA technique is performed by filtering the cepstrum to remove any fast changing spectrum values.

Simultaneous masking describes the situations where the masker is present for the whole time that the signal occurs. Based on this idea, we compute a masking threshold (masker) as a function of frequency for a given speech frame from its power spectrum. Figure 1(a) shows the power spectrum of a clean, speech frame with the masker present. Frequency components below the masking threshold, are components that are not heard by the human auditory system due to masking effects. These frequency components are replaced with the masking threshold resulting in a magnitude spectrum such as the one shown in Figure 1(b).

The performance of speech recognition systems degrades when a mismatch between the training and testing environments exists. If a speech recognition system is trained on a clean environment and is then utilised in a noisy environment, it will perform badly due to the mismatched conditions [15].

Adding white noise to the signal results in an constant addition to the log power spectrum. If we take any given point in the power spectrum, we can say that it has a given SNR value. That is, the SNR values will be smaller in the troughs of the power spectrum than in the peaks. If we can ignore or threshold the troughs of the power spectrum based on some auditory property, then we can obtain a spectrum which is robust to additive noise.

We use this simultaneous front-end in both the training data and in the front-end during the recognition. Therefore, the environmental mismatch is reduced due to the feature estimates being robust to noise.

The simultaneous masking procedure presented in this paper is described by the following. Firstly, the power spectrum of a speech frame is computed using a fast Fourier transform algorithm. The masking threshold is then calculated from the power spectrum as a function of frequency. In these experiments, we have calculated the masking threshold by using the critical band masking curve used in the PLP algorithm [3]. This is the same as the Mel filter bank anaylsis except for the shape of the critical band curve. The portions of power spectrum which are lower than the masking threshold are discarded and replaced by the corresponding portions from the masking threshold spectrum.

This can be stated as follows. Let $P(f)$ be the power spectrum of the speech frame computed through FFT algorithm and $M(f)$ the masking threshold spectrum obtained from $P(f)$ by using simultaneous masking. As mentioned, the portions of $P(f)$ lower than $M(f)$ are not heard due to masking effects and hence can be discarded. The modified power spectrum $\hat{P}(f)$ which takes care of these masking effects is then computed as follows:

$$\hat{P}(f) = \max[P(f), M(f)],$$

where f is frequency in the range 0 to 4000 Hz for the speech signal sampled at 8000 Hz.

Cepstral features from the modified power spectrum $\hat{P}(f)$ are then calculated using either a homomorphic or the linear prediction analysis technique. The experiments presented in this paper use the front-ends FFTHCC and FFTLCC for the homomorphic and LP analysis respectiviely as described in [5]. We also apply the masker to other front-ends such as a uniform filter bank LP-based front-end (UNILCC) and a Mel filter bank LP-based front-end (MELLCC) to observe their effect on recognition performance.

3. RECOGNITION AND DATABASE

Speech recognition experiments performed in this paper use HMM based isolated word and continuous speech recognisers. For the isolated word recognition expeirments, we use the ISOLET database. This database consists of the 26 letters of the english alphabet. There are 4680 utterances from 90 speakers used for training of 26 HMM models and 1560 utterances from 30 speakers used for testing. The database consists of the same number of males as females and different speakers are used for testing and training.

For continuous recognition, we use the TIMIT speaker-independent, continuous speech recognition database to train 48 context-independent models as proposed in [16]. We only use the SX and SI sentences leaving 8 utterances per speaker. Therefore, our experiments use 3696 training sentences from 462 speakers of which 326 are male. Testing is performed using 1344 testing sentences from 168 speakers of which 112 are male.

Both these databases are sampled at 16 kHz with a 16 bit resolution. We have decimated them to 8 kHz using a low pass filter with a cutoff frequency of 3.5 kHz. From this, speech frames are computed every 15 ms with a 45 ms hamming window. We compute 10th order cepstral features using either the homomorphic and LP based front-ends with and without masking.

Both the continuous and isolated word recongition models use simple left-to-right HMM models. The continuous speech models contain 3 states except for the closures and silences (epi, sil, cl, vcl) which are modelled using only a single state. The isolated word models are modelled with 5 states and both recognition systems use 5 Guassian mixtures in all states with diagonal covariances.

[1] Note that the auditory masking properties have been successfully used in the past for improving the performance of speech and audio coders [8, 9, 10, 11, 12].

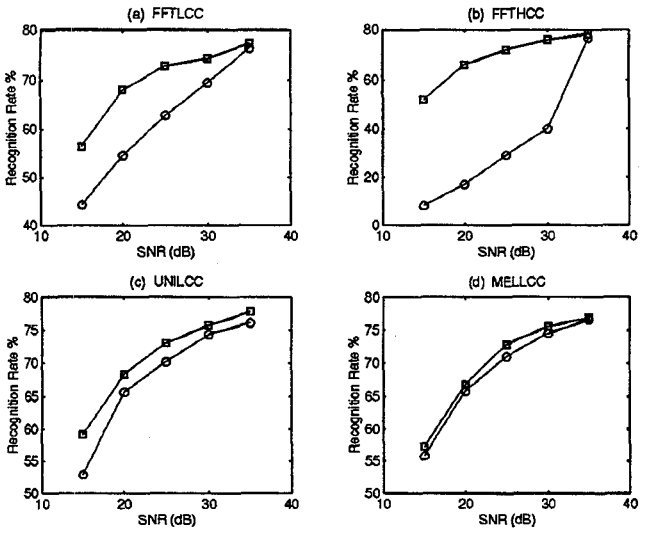

Figure 2. Recognition performance of using no masking (circles) and using masking (squares) for four different front-ends with isolated word modelling.

All models are trained on speech using the various front-ends with and without simultaneous masking. For studying the effect of additive noise distortion, white Gaussian noise is added to the speech signal and speech recognition performance is investigated as a function of signal-to-noise ratio (SNR). The effect of channel distortion is evaluated by filtering the test signal using the same filter as presented in [5]. In the second set of experiments, we use the ISOLET database to show results using RASTA [6] to simulate a more complete model for the masking property of the human ear.

4. RESULTS

4.1. Experiment 1 - Simultaneous Masking

Figure 2(a) shows the results using the FFTLCC front-end with and without masking in the presence of white noise. Firstly, we notice that the recognition performance is not degraded for clean speech even though Figure 1 showed that some frequency peaks are removed from the spectrum due to masking. Compared to the FFTLCC front-end, the FFTLCC based masking front-end shows an improvement at all SNR values. In terms of SNR the masking front-end shows an improvement of more than 5 dB.

Results showing the recognition performance of the homomorphic based FFT front-end (FFTHCC) are shown in Figure 2(b). The FFTHCC front-end on its own performs poorly in the presence of white noise. Using the masking based FFTHCC front-end, improvement in performance in terms of SNR is equivalent to over 15 dB.

These results show that the auditory property of simultaneous masking has a significant effect on speech recognition performance in the presence of white noise. Simultaneous masking removes the parts of the frequency spectrum that are most effected by noise. We compare this to a warped filter bank front-end that averages the frequency spectrum for each filter bank and is thus robust to addititive noise distortions.

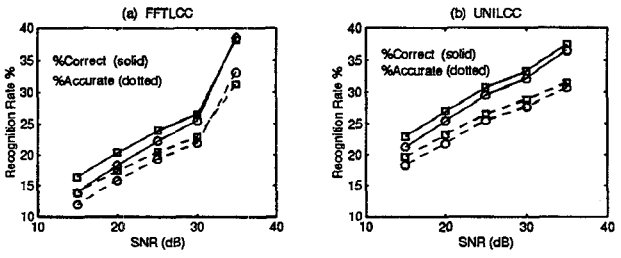

Figure 3. Recognition performance using masking for continuous speech recognition in white noise (a) using the FFTLCC front-end and (b) using the UNILCC front-end.

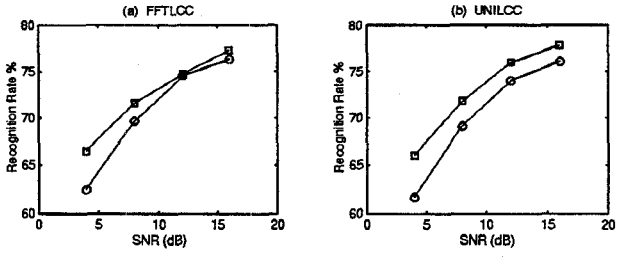

Figure 4. Recognition performance using masking for isolated word recognition in channel distorted speech (a) using the FFTLCC front-end and (b) using the UNILCC front-end.

We compare the FFT based front-ends using the simultaneous masking procedure with uniform (UNILCC) and Mel-spaced (MELLCC) LP-based filter bank front-ends. Figures 2(c) and (d) show the recognition performance with and without masking using the UNILCC and MELLCC front-ends respectively. From these results, we can see that the UNILCC front-end with masking outperforms the MELLCC front-end with masking. We notice that the masking procedure does not improve the performance of the MELLCC front-end significantly.

As found in the isolated word recognition case, masking improves the performance of a continuous speech recogniser in the presence of white noise. Figures 3(a) and (b) show the improvement in recognition performance of the FFTLCC and UNILCC front-ends using masking. Again, the masking based front-ends outperform the standard front-ends at most SNR values. Not shown are the results for the FFTHCC and MELLCC front-ends. The effect of continuous speech recognition for these front-ends is similar to the isolated recognition case.

To complete the study of this masking technique in noise, we show the recognition performance using simultaneous masking in the presence of a channel distortion. Figures 4(a) and (b) show the recognition performance using masking on the FFTLCC and UNILCC front-ends for isolated word recognition. These results show that the masking front-end also improves the recognition performance in the presence of a distorted channel.

This confirms that the masking front-end improves the performance of a recognition system for isolated word and continuous speech recognition systems in the presence of white noise and channel distortions.

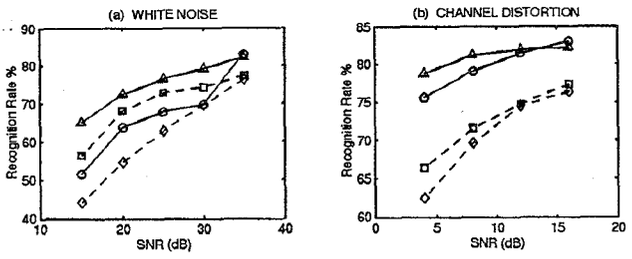

Figure 5. Recognition performance using masking and RASTA for isolated word recognition in (a) white noise and (b) in the presence of channel distorted speech. (dotted diamond - no processing, dotted square - simultaneous masking, solid square - RASTA, solid traingle - masking + RASTA)

4.2. Experiment 2 - Using simultaneous and forward masking

To complete the experiments presented in this paper, we perform RASTA processing with this masking front-end. By combining the two techniques, this front-end now models a more complete auditory masking model as it takes into account the properties of simultaneous and non-simultaneous masking.

Figure 5(a) shows the recognition performance using the FFTLCC front-end for three cases: using simultaneous masking, using RASTA and combining the two techniques. This figure shows that the recognition performance increases by adding the non-simultaneous property to the front-end. In terms of SNR, RASTA processing adds almost an additional 5 dB of improvement over the masking front-end alone. Combining both the RASTA technique and the property of simultaneous masking, a robust front-end with better than a 10 dB equivalent SNR improvement is found compared to the FFTLCC front-end.

For the channel distortion case, the recognition performance also increases due to the addition of RASTA processing as shown in Figure 5(b).

5. CONCLUSION

We have shown a robust front-end for speech recognition based on the auditory property of simultaneous masking. This front-end improves the performance in terms of SNR more than 5dB for an isolated word HMM recogniser and 3 dB for a continuous speech HMM recogniser. We found that by incorporating a uniform filter bank with the masking front-end, additional robustness was seen.

Adding the simultaneous masking technique to a technique modelling the non-simultaneous masking property such as RASTA, we were able to show further improvements in recognition performance at lower SNR levels.

REFERENCES

[1] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. 28, No. 4, pp. 357-366, Aug. 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, Vol. 87, No. 4, pp. 1738-1752, Apr. 1990.

[4] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environments", *Computer Language and Speech*, Vol. 1, pp. 109-130, 1986.

[5] K.K Paliwal and B.S. Atal, "A comparative study of feature representations for robust speech recognition in adverse environments", *Proc. Int. Conf. Spoken Language Processing*, Yokohama, Japan, Sept. 1994.

[6] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 4, pp. 578-589, Oct. 1994.

[7] K. Aikawa, H. Singer, H. Kawahara and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp 668-671, 1993.

[8] E. Zwicker and U.T. Zwicker, "Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system", *J. Audio Eng. Soc.*, Vol. 39, No. 3, pp. 115-125, Mar. 1991.

[9] M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 937-940, 1985.

[10] D. Sen and W.H. Holmes, "Perceptual enhancement of CELP speech coders", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. II, pp. 105-108, 1994.

[11] G. Theile, G. Stoll and M. Link, "Low bit-rate coding of high quality audio signals: An introduction to the MASCAM system", *EBU Review – Technical*, No. 230, Aug. 1988.

[12] J.D. Johnston and K. Brandenburg, "Wideband coding – Perceptual considerations for speech and music", in S. Furui and M.M. Sondhi (Eds.), *Advances in Speech Signal Processing*, Marcel Dekker, New York, pp. 109-140, 1992.

[13] B. Scharf, "Critical bands", in J. Tobias (Ed.), *Foundations of Modern Auditory Theory*, Academic, New York, pp. 159-202, 1970.

[14] E. Zwicker, "Masking and psychological excitation as consequences of ear's frequency analysis", in R. Plomp and G.F. Smoorenburg (Eds.), Sijthoff, Leyden, The Netherlands, 1970.

[15] B.H. Juang, "Speech recognition in adverse environments" Computer Speech and Language, Vol. 5, pp. 275-294, 1991.

[16] K. Lee and H. Hon, "Speaker-Independent phone recognition using Hidden Markov Models", *IEEE Trans. Speech and Audio Processing*, Vol. 37, No. 11, pp 1641-1648, Nov. 1989.