International Symposium on Signal Processing and its Applications, ISSPA, Gold Coast, Australia, 25-30 August, 1996.
Organized by the Signal Processing Research Centre, QUT, Brisbane, Australia.

THE EFFECT OF EXTERNAL AND MIDDLE EAR FILTERING ON AUTOMATIC PHONEME RECOGNITION

Beng T. Tan
Dept. of Elect. and Comp. Eng.,
The University of Newcastle,
NSW 2308, Australia.

Phillip Dermody
National Acoustic Laboratories,
126 Greville Street,
NSW 2067, Australia.

Minyue Fu, Andrew Spray
Dept. of Elect. and Comp. Eng.,
The University of Newcastle,
NSW 2308, Australia.

ABSTRACT

This study investigates the effects of using different external and middle ear models as the preemphasis filter for speech recognition. We find that an external and middle ear model derived from physiological data of human auditory systems gives the highest recognition rate. This auditory based preemphasis filter also gives good a recognition rate for the consonant sounds (fricatives, stops, liquids and diphthongs).

1. INTRODUCTION

It is common to preemphasize speech signals with a linear filter, $1 - az^{-1}$, before speech is parameterized. This filter is motivated by the fact that voiced speech carries an approximate -6dB/octave slope in the spectrum on average. The need for preemphasis in speech recognition has been justified in many recognition systems.

It is well-known that the external ear and middle ear play an important role in the sensitivity of our hearing with respect to different frequencies. Naturally, an alternative preemphasis method for speech recognition is to use an auditory based filter derived from external and/or middle ear models. However, there seems no reported result on the use of these ear models in preemphasis for speech recognition.

Our purpose here is to investigate the effect of external and middle ear filtering on phoneme recognition.

2. EXTERNAL EAR AND MIDDLE EAR

The sound wave perceived by the external ear gets amplified and filtered. If we assume that the sound wave is incident at an angle of 45° relative to the front-back axis of the human head, the approximated external ear transfer function is shown by the solid line in Fig. 1(a). The impedance matching process of the middle ear provides amplification of sound pressure during the transmission from the tympanic membrane to oval window. The frequency characteristics of the middle ear are found [6] by measuring the round window volume velocity for a constant sound pressure at the tympanic membrane as shown in Fig. 1(b) (dashdot line). Other external and middle ear functions which are adopted in different auditory models are also shown in Fig. 1.

The composite transfer function of external [8] and middle [6] ear is shown in Fig. 2 (dashed line). Hewitt *et al.* [3] approximate the composite filter by a bandpass function as shown in Fig. 2 (solid line). Seneff [7] and Van Immerseel *et al.* [10] ignore the middle ear transfer function in their auditory models whereas Kates [5] and Deng *et al.* [1] ignore the external ear transfer function. All these models will be tested.

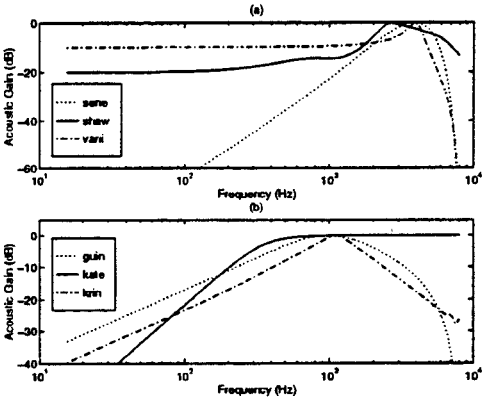

Figure 1: (a) 'shaw' (solid line) is the human external ear transfer function given by (Shaw, 1974). 'sene' (dotted line) and 'vani' (dashdot line) are the external ear models which are adopted in the auditory models of Seneff (1988) and Van Immerseel *et al.* (1992), respectively. (b) 'krin' (dashdot line) and 'guin' (dotted line) are the middle ear transfer functions (round window volume velocity per unit pressure) of human (Kringlebotn, 1985) and cats (Guinan *et al.*, 1967), respectively. 'kate' (solid line) is the acoustic gain of the middle ear adopted by Kates (1991).

3. PHONEME RECOGNITION ENVIRONMENT

Our phoneme recognition tests are evaluated on the prototype version (1988) of the TIMIT database. We use the DR1 region only. The training tokens consist of 12 females and 17 males, and the testing tokens consist of 3 females and 5 males. There are 290 sentences for training and 80 sentences for testing in total. Seven groups of allophones are identified and within-group confusions are not counted as errors [9]. The speech signals are sampled at 16 kHz.

Each analysis frame has a duration of 20 ms with a 10 ms overlap. After the speech is preemphasized, 12 Mel-frequency cepstral coefficients per frame are computed.

The phoneme recognizer consists of 59 phone models. Each phone is modeled by a three state left-to-right HMM. The output probability distribution of each state is modeled

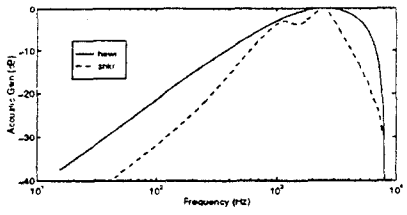

Figure 2: 'shkr' (dashed line) is the composite transfer function of human external (Shaw, 1974) and middle (Kringlebotn *et al.*, 1985) ear. 'hewi' (solid line) shows the transfer function used by Hewitt *et al.* (1992).

by a mixture of three multivariate Gaussian density functions with a diagonal covariance matrix. HMM parameters are initialized with the segmental K-mean algorithm and estimated using the Baum-Weltch re-estimation algorithm.

4. RESULTS AND DISCUSSION

The phoneme recognition results with different preemphasis filters are shown in Fig. 3. Now we offer some discussion of these results:

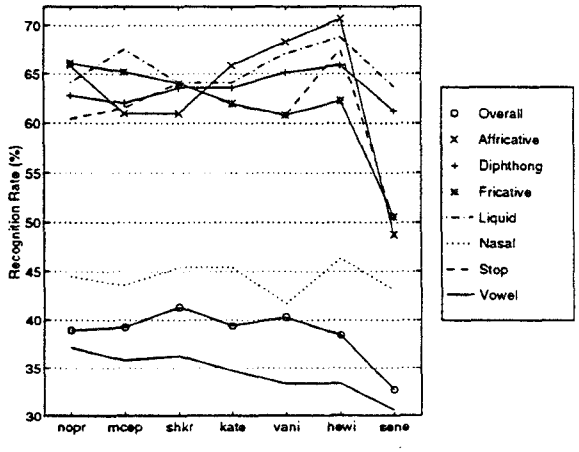

Figure 3: The recognition rate of untrained samples. 'mcep' uses the conventional preemphasis filter, $1 - az^{-1}$, with preemphasis factor of 0.95. No preemphasis filter is used in 'nopr'. 'shkr' uses the composite function of external ear (Shaw, 1980) and middle ear function (Kringlebotn, 1985). 'kate', 'sene', 'vani' and 'hewi' are based on the auditory models by Kates (1991), Seneff (1988), Van Immerseel *et al.* (1992) and Hewitt *et al.* (1992), respectively.

1) The results suggest that the choice of preemphasis filter does affect the performance of a speech recognizer. The preemphasis filter 'shkr' derived from the physiological data of external [8] and middle ear [6] gives the highest phoneme recognition rate. We believe that this is because its transfer function has most weight in the frequency range of 500 Hz to 5000 Hz. An interesting property of this model is that it also gives a good recognition rate (64%) for the consonant sounds (fricatives, stops, liquids and diphthongs).

2) 'sene' yields the lowest overall phoneme recognition rate. We suspect that this is caused by the steep slope (about $+55dB/dec$) at the frequency below 4 kHz. The recognition rate of affricatives, fricatives and stops are particularly low. Note that the spectral energy of these sounds is concentrated in the range from 500 Hz to 4000 Hz [4].

3) It is interesting to note that the recognition rate of both 'vani' and 'kate' are close. A low-pass filter is used in 'vani' [10] and a high-pass filter is used in 'kate' [5]. This is not surprising because experiments of speech intelligibility [2] shows that speech is highly perceptible when heard through a low-pass filter with high a cut-off frequency or a high-pass filter with a low cut-off frequency. The conventional preemphasis, 'mcep', gives a fair recognition rate and is compatible to the result of 'vani'and 'kate'. Also, the case without preemphasis ('nopr') performs quite well too.

4) In comparision with the conventional preemphasis ('mcep') and the case without preemphasis ('nopr'), auditory based preemphasis filters seem to have some edge, although the difference is not very significant. We note, however, that the tests are done in a noise-free enviornment, i.e. the database is recorded without noises. Further tests need to be done to study the effect of auditory based preemphasis filters in adverse environments.

5. REFERENCES

[1] L. Deng, C. Geisler, and S. Greenberg. A composite model of the auditory periphery for the processing of speech. *J. Phonetics*, 16:93–108, 1988.

[2] N. French and J. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Amer.*, 19(1):90–119, 1947.

[3] M. Hewitt, R. Meddis, and T. Shackleton. A computer model of a cochlear-nucleus stellate cell: Responses to amplitude-modulated and pure-tone stimuli. *J. Acoust. Soc. Amer.*, 91(4):2096–2109, 1992.

[4] G. Hughes and M. Halle. Spectral properties of fricative consonants. *J. Acoust. Soc. Amer.*, 28:303–310, 1956.

[5] J. Kates. A time-domain digital cochlear model. *IEEE Trans. Signal Processing*, 39(12):2573–2592, 1991.

[6] M. Kringlebotn and T. Gundersen. Frequency characteristics of the middle ear. *J. Acoust. Soc. Amer.*, 77(1):159–164, 1985.

[7] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *J. of Phonetics*, 16:55–76, 1988.

[8] E. Shaw. The external ear. In W. Keidel and W. Neff, editors, *Handbook of sensory physiology*, volume 5/1, pages 455–490. Springer, Berlin, 1974.

[9] B. T. Tan, M. Fu, and P. Dermody. The use of wavelet transforms in phoneme recognition. In *To be appeared in Proc. ICSLP-96*, 1996.

[10] L. M. Van-Immerseel and J.-P. Martens. Pitch and voiced/unvoiced determination with an auditory model. *J. Acoust. Soc. Amer.*, 91(6):3511–3526, 1992.