EFFECT OF FILTER BANDWIDTH AND SPECTRAL SAMPLING RATE OF ANALYSIS FILTERBANK ON AUTOMATIC PHONEME RECOGNITION

Feipeng Li[1], Hynek Hermansky[1,2]

[1]Center for Language and Speech Processing
[2] Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD, 21218
fli12@jhmi.edu, hynek@jhu.edu

ABSTRACT

In this study we investigate the effect of filter bandwidth and spectral sampling rate of analysis filterbank for speech recognition. Two experiments are conducted to evaluate the performance of an automatic phoneme recognition system on clean speech and speech in noise as the filter bandwidth increases from 0.5 to 3.5 ERB and the spectral resolution changes from 1, 1.5, 2, 3, 4, to 6 samples per Bark. Results indicate that the optimum filter bandwidth varies for different speech sounds at different frequency ranges. A spectral sampling of 4 filters per Bark with the filter bandwidth being ≈ 1 ERB produces the best performance on average.

Index Terms: filter bandwidth, spectral resolution, phoneme recognition

1. INTRODUCTION

Speech analysis is an indispensable process for automatic speech recognition. An optimal filterbank suppresses interfering noise while maximizes the speech information being extracted. Filter bandwidth and spectral sampling rate (i.e., number of filters/Bark) are the two key parameters of an analysis filterbank.

Past studies show that auditory filter bandwidth has an important effect on human speech perception, especially in adverse environments. Comparative study across various biological species indicate that the auditory filter of human cochlea is considerably sharper than that of other mammals, which may facilitate speech communication [6]. People with hearing loss, characterized by an abnormally wide psychoacoustic tuning curve (i.e., reduced frequency selectivity), generally have great difficulty understanding speech in noise [8]. In [2] the effect of reduced spectral resolution on speech perception was simulated by smoothing the envelope of the squared short-time Fourier transform by convolving it with a Gaussian-shaped filter. It was shown that the speech reception threshold (SRT) level, defined as the signal-to-noise ratio (SNR) at which subjects can understand 50% of spoken words, elevated as the size of smoothing window approximates the critical bandwidth. In another study, [5] simulated the loss of frequency selectivity in people with moderate to severe cochlear hearing loss on normal hearing listeners. While the intelligibility of speech in quiet was hardly affected by spectral smearing, even with a broadening factor of 6, the impact of reduced frequency selectivity on speech intelligibility in noise is substantial.

A common measure of human auditory critical bandwidth is the Equivalent Rectangular Bandwidth (ERB)[8], defined as the bandwidth of a rectangular window that has the same amount of masking in noise.

$$\text{ERB} = 0.108 \times f + 24.7 \tag{1}$$

where ERB is the bandwidth in Hz, and f is the center frequency in kHz. Another popular measure of critical bandwidth is the Bark scale, derived based on the subject measurements of loudness [7].

$$\text{Bark} = 6\,\text{asinh}(f/600) \tag{2}$$

A Bark is about 2.86 times the size of ERB at 0.1 kHz. It then keeps decreasing until it hits 1.57 kHz, beyond which the Bark-to-ERB ratio remains close to 1.5. Both ERB and Bark scale were derived based on the results of human psychoacoustic experiments.

Recently, it was shown that auditory filters optimized for a combined set of vowels and consonants based on the criterion of band independence and maximum mutual information matches the human physiological data [10]. Moreover, the optimal filter derived based on the average spectrotemporal dependencies of continuous speech closely matches the critical bandwidth of human auditory system [11].

Another important parameter for speech analysis is the spectral sampling rate, i.e., number of filters per critical band.

This work was supported in parts by the DARPA RATS project D10PC0015, IARPA BABEL project W911NF12-C-0013, and by the Johns Hopkins Center of Excellence in Human Language Technologies. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA, IARPA or JHU HLTCOE.

Shannon and his colleagues investigated the recognition of clean speech with only 1 to 4 frequency channels [3]. The spectral detail information was removed by multiplying the subband envelopes with band-limited noise. Results show that the average percent correctness of consonants and vowels is around 90% with only 4 channels.

To summarize, past studies indicate that filter bandwidth and spectral sampling rate have important effect on human speech perception. In this study, we investigate the effect of these two factors on automatic phoneme recognition in both clean and noisy conditions with an aim to find an optimal combination of the two parameters for speech analysis.

2. SYSTEM DESCRIPTION

The automatic phoneme recognition system takes an ANN/HMM hybrid architecture [1]. The speech signal is classified into 40 phonemes by a three-layer multi-layer perceptron (MLP), followed by a Viterbi decoder which produces a mostly likely phone sequence based on the posterior probability produced by the MLP. The speech recognizer is trained on TIMIT speech corpus, which contains 4 hours of high-quality speech, produced by a larger number of talkers. The speech signal is sampled at 16kHz.

The power spectra of TIMIT speech increases by 30 dB from 0.1 to 0.3 kHz, where it reaches the peak, and remains at the same level until 0.7 kHz, then it drops to 25 dB at 8 kHz with a slope of 12 dB/octave. Two types of noises with different spectral shape, babble and subway, are used to mask the speech signal. Assuming that the speech and noise have the same amount of power, the power spectra of babble noise is very similar to that of TIMIT speech, except for slight amount of differences in the low frequency range below 0.2 kHz and high frequency range above 4 kHz. In contrast, the power of subway noise is concentrated around 0.15 to 0.3 kHz, where it is about 20 dB stronger than that of TIMIT speech. Beyond 0.4 kHz, the intensity of TIMIT speech is consistently higher than the subway noise by about 20 to 40 dB. Since speech information is distributed mainly in the mid-frequency range from 0.3 to 3 kHz, the babble noise introduces much more masking effect than the subway noise.

The speech signal is encoded by the frequency domain linear prediction modulation (FDLPm) feature [9], which provides a parametric representation of the Hilbert envelope of subband signal. The context window and maximum frequency of temporal modulation are chosen to be 300 ms and 32 Hz respectively. Given the number of filters and bandwidth, the frequency decomposition is implemented by first calculating the frequency spectrum of speech signal using the DCT transform, then the speech spectrum is divided into multiple bands by multiplying the DCT coefficients with a set of window functions of specified bandwidth, which is described by

$$w_k = 1 - 0.5(1 - \cos(0.729\,\pi(f - f_k)/B) \qquad (3)$$

where f and f_k are the frequency variable and center frequency of the k^{th} band respectively, both in Bark scale; B is the bandwidth of the window, defined as the frequency range where the window amplitude is greater than the 3 dB threshold (0.707).

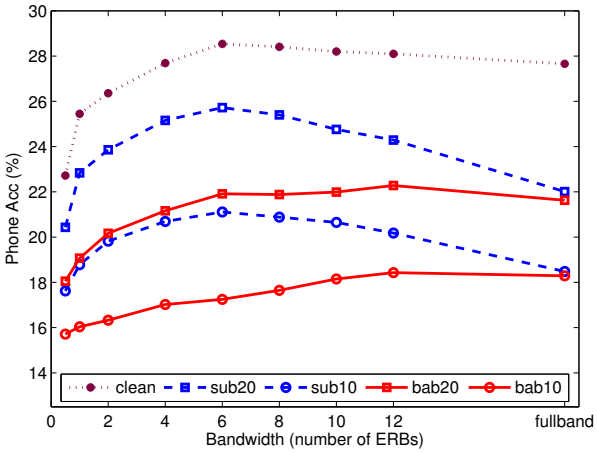


Fig. 1. Phone Accuracy of the 10th bands (cf = 1247 Hz) as a function of bandwidth in clean and noisy conditions. "sub10" and "bab10" refer to subway/babble noise at 10 dB SNR.

3. EXPERIMENTS

A pilot study is conducted to explore how speech recognition is dependent on the bandwidth of a single auditory filter (window), with the center frequency being fixed at 1247 Hz, which is close to the center of the full frequency range on Bark scale. The filter bandwidth varies from 0.5, 1, 2, 4, 6, 8, 10, 12 ERB to fullband (i.e., no filtering). Results (refer to Fig. 1) indicate that the phone accuracy increases sharply from 0.5 to 1 ERB, then it becomes flat but keeps increasing slowly until it reaches the maximum at 6 ERB, suggesting that 1 ERB might be the critical bandwidth.

Two experiments are conducted to evaluate the effect of filter bandwidth and spectral sampling rate on the performance of an automatic phoneme recognizer. The first experiment aims to explore how phoneme recognition is dependent on the bandwidth of individual auditory filters. The second experiment aims to determine the optimum combination of filter bandwidth and spectral sampling rate. Both experiments are conducted under clean and noisy conditions.

3.1. Experiment I

The first experiment aims to check whether the "critical" bandwidth of 1 ERB, identified from the pilot study, applies to other frequency ranges as well. Three bands (7, 13, and 17th bands), with center frequencies of 691, 2111, and 4132 Hz

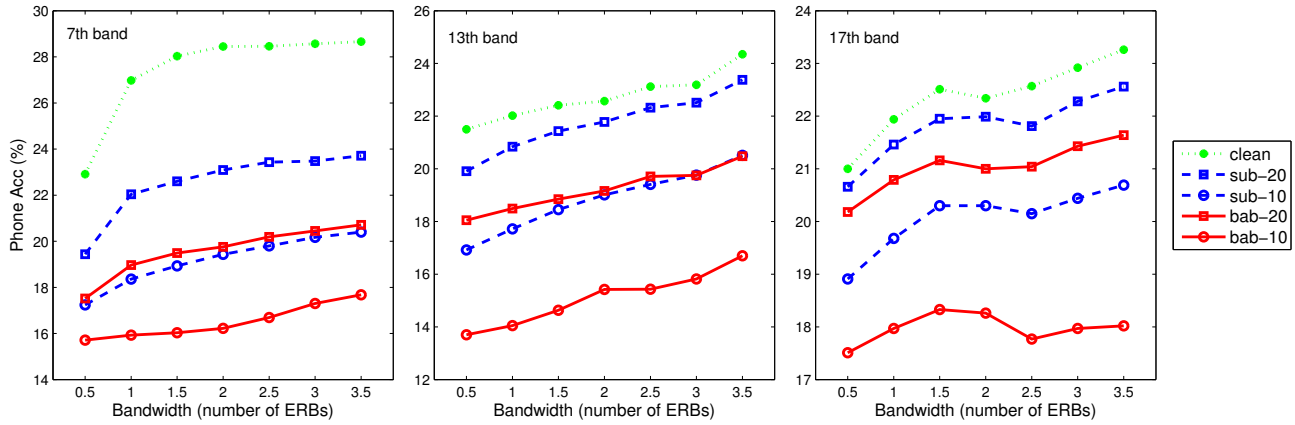
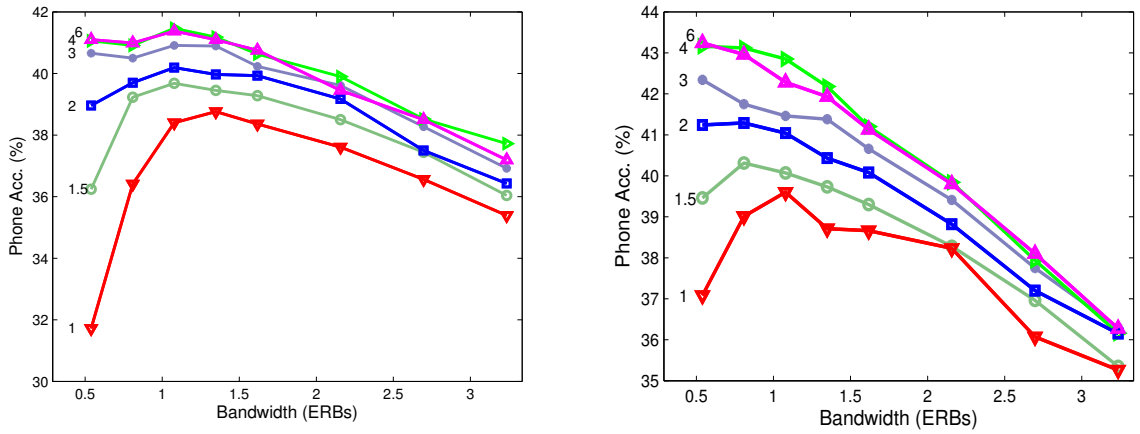


Fig. 2. Phone accuracy of the 7, 13, and 17^{th} band (cf = 691, 2111, 4132 Hz respectively) as a function of bandwidth



(a) f=(560, 1278)Hz (b) f=(1532, 3065)Hz

Fig. 3. The optimum combination of bandwidth and spectral sampling rate (labeled on the left side of each curve) for phoneme recognition.

respectively, are selected for the experiment. The filter bandwidth changes from 0.5 to 3.5 ERB with a step size of 0.5 ERB. The ASR system is tested under both clean and noisy conditions with babble and subway noise at 20 and 10 dB SNR.

Results (refer to Fig. 2) show that the phone accuracy of the 7^{th} band increases relatively fast from 0.5 to 1 ERB. It then slows down at 1.5 ERB and becomes nearly flat, suggesting that 1 ERB is critical for phoneme recognition in low frequency range. In contrast, the 13^{th} and 17^{th} bands show no critical changing in slope at 1 ERB. The gap between clean speech and noisy speech is much bigger for the 7^{th} band than that of the 13^{th} and 17^{th} bands, where the power spectra of the speech is significantly larger than the masking noise, suggesting that filtering is important for the rejection of masking noise in certain frequency bands. Since most of the speech energy is located from 0.3 to 1.5 kHz, it is important that the minimal bandwidth being no less than 1 ERB.

3.2. Experiment II

The second experiment aims to determine the optimum combination of filter bandwidth and spectral sampling rate (i.e., number of filters per Bark) for different frequency range. Two frequency ranges: [560,1278] Hz and [1532, 3065] Hz, each covers 4 Barks along the auditory frequency axis, are selected for the experiment. The spectral sampling rate increases from 1, 1.5, 2, 3, 4 to 6 filter/Bark, while the 6dB filter bandwidth (i.e., threshold=0.5) changes from 0.5 to 3 Bark with a step size of 0.5 (Bark scale). [1].

Results (refer to Fig. 3) show that the performance of the phoneme recognition system increases as the overlap between neighboring filters increases from less than 25% (1 fil-

[1] A 6dB bandwidth in Bark scale $\approx$ 3dB bandwidth in ERB $\times$ 1.0785

ter/Bark) to 75% (4 filters/Bark) for both frequency regions. Further increasing in spectral sampling rate does not help in improving the performance. The results in noisy conditions are very similar to the results in clean condition. In addition, the phoneme accuracy reaches the maximum when the bandwidth $\approx$ 1 ERB for frequency range (560, 1278)Hz and $\approx$ 0.5 ERB for frequency range (1532, 3065)Hz respectively, suggesting that the optimum bandwidth may change for different frequency ranges.

4. SUMMARY

In this study, we investigated the effect of two key parameters for speech analysis, i.e., filter bandwidth and spectral sampling rate, on phoneme recognition under clean and noisy conditions. Two independent experiments are conducted to identify the best filter bandwidth for single filters and optimum combination of filter bandwidth and spectral sampling rate for filterbank. Results show that the performance of phoneme recognition system is highly dependent on the two parameters. A filter bandwidth of 1 ERB produces the best average phone accuracy under both clean and noisy conditions. A sampling rate of 4 filters per Bark (i.e., 75% overlap) produces the best performance for phoneme recognition.

5. REFERENCES

[1] Bourlard, H. and Morgan, N., "Connectionist Speech Recognition – a hybrid approach.", Kluwer Academic Publishers, Boston. 1994.

[2] M. ter Keurs, J. M. Festen, and R. Plomp, "Effect of spectral envelope smearing on speech reception. I," J. Acoust. Soc. Am., 91(5):2872-2880, 1992.

[3] R. V. Shannon, F-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science, 270:303–304, 1995.

[4] Q. J. Fu, R. V. Shannon, and X. Wang, "Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing," J. Acoust. Soc. Am., 104(6):3586–96, 1998.

[5] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise ," J. Acoust. Soc. Am., 94(3): 1229–1241, 1993.

[6] C. Shera, J. J. Guinan, and A. J. Oxenham, "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," Proc. Natl. Acad. Sci., 99(5):3318–23, 2002.

[7] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception," in *Frontiers of Speech Communication*, New York: Academic, 1979.

[8] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data.," Hearing research, 47(1–2):103-38, 1990.

[9] Ganapathy, S., Thomas, S., and Hermansky, H., "Temporal envelope compensation for robust phoneme recognition using modulation spectrum.", J. Acoust. Soc. Amer. 128(6):3769–3780, 2010.

[10] Lewicki, M. S., "Efficient coding of natural sounds.", Nature Neuroscience. 5(4):356–363, 2002.

[11] Rasanen, O., "Average spectrotemporal structure of continuous speech matches with the frequency resolution of human hearing.", Proc. Interspeech, 2012.