Speech-Signal-Based Frequency Warping

Kuldip Paliwal, *Member, IEEE*, Benjamin Shannon, James Lyons, and Kamil Wójcicki

***Abstract*—The speech signal is used for transmission of linguistic information. High energy portions of the speech spectrum have higher signal-to-noise ratios than the low energy portions. As a result, these regions are more robust to noise. Since the speech signal is known to be very robust to noise, it is expected that the high energy regions of the speech spectrum carry the majority of the linguistic information. This letter tries to derive a frequency warping function directly from the speech signal by sampling the frequency axis non-uniformly with the high energy regions sampled more densely than the low energy regions. To achieve this, an ensemble average short-time power spectrum is computed from a large speech corpus. The speech-signal-based frequency warping is obtained by considering equal area portions of the log spectrum. The proposed frequency warping is shown to be similar to the frequency scales obtained through psycho-acoustic experiments, namely the mel and bark scales. The warping is then used in filterbank design for automatic speech recognition experiments. The results of these experiments show that cepstral features based on the proposed warping achieve performance under clean conditions comparable to that of mel-frequency cepstral coefficients, while outperforming them under noisy conditions.**

***Index Terms*—Mel scale, bark scale, speech-signal-based frequency warping, speech-signal-based frequency cepstral coefficient (SFCC), robust automatic speech recognition (ASR).**

I. Introduction

BETWEEN HUMAN auditory and speech production systems, it is believed that the auditory system came first. As a result, we expect that the production system has evolved over time to match the auditory system. The production system generates the speech signal, which gets processed by the auditory system. It is the acoustic speech signal which mediates between these systems. Thus, it is only natural to expect that the properties of the acoustic signal can tell us about both the human speech production system and the human auditory system (for some of the references that show this link, see [1]–[6]). The primary objective of this work is to find a frequency warping function based purely on the properties of the acoustic speech signal. This warping function can then be compared with the well-established auditory based scales. This comparison will tell us how well the production and auditory systems are matched.

In the past, two very similar frequency warping scales have been derived through psycho-acoustic studies. These are the mel scale [7] and the bark scale [8]. These warping functions are very popular in the speech processing literature and have been employed in many speech processing applications [9].

Manuscript received October 13, 2008; revised December 11, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Philip Loizou.

The authors are with the Signal Processing Laboratory, Griffith University, Nathan QLD 4111, Australia (email: k.paliwal@griffith.edu.au, poboxben@gmail.com, j.lyons@griffith.edu.au, k.wojcicki@griffith.edu.au).

Digital Object Identifier ??.????/LSP.200?.??

In addition to these psycho-acoustic approaches, data-driven or speech-signal-based approaches have also been proposed in the literature to derive frequency warping function with the performance of automatic speech recognition (ASR) in mind. For example, Burget and Hermansky [10] have employed linear discriminant analysis to maximise the separability between linguistic classes, while minimising their within-class scatter. In another study, Biem *et al.* [11], [12] have used a minimum classification error training algorithm to derive a frequency warping function that minimises the ASR classification error. Data-driven approaches have also been developed to show that the warping function obtained by making similar enunciations of different speakers as translations in the speech spectra is similar to the mel function [13]–[15].

In the present work, we derive a frequency warping directly from the acoustic speech signal using a simple approach based on the following arguments. The high energy (formant) regions of the speech spectrum have higher signal-to-noise ratios (SNRs) than the low energy regions, which makes them more robust to noise. Since the speech signal is known to be very robust to noise, it is expected that these high energy portions of the speech spectrum carry the majority of the linguistic information. It is logical to perform finer sampling of the speech spectrum regions which contain more linguistic information; *i.e.*, of the high energy regions. To achieve this an ensemble average short-time power spectrum, $\bar{P}(f)$, is computed over a large speech corpus. The frequency axis is then divided into M non-overlapping intervals, such that area under $\log \bar{P}(f)$ curve for each interval is the same. A mapping from frequency axis to warped frequency axis is obtained by taking the middle frequency of each interval. We show that the resulting warping function is similar to the mel and bark scales.

Frequency warping has been employed in speech processing to derive features for ASR with much success. For example, the mel-frequency cepstral coefficients (MFCCs) have become the de-facto standard for the current day ASR. The second objective of this work is to compare the ASR performance of cepstral features based on the proposed frequency warping with the traditional MFCCs. For this purpose, ASR experiments are conducted on the TIMIT speech corpus. The results of these experiments show that the proposed frequency warping achieves ASR performance comparable to that of mel-frequency warping under clean conditions. In addition, we show that the cepstral features based on the proposed warping are more robust to noise than MFCCs.

The rest of this letter is organized as follows. Section II describes the speech-signal-based frequency warping procedure. Section III details our experiments along with their results. The discussion and future directions are presented in Section IV.

discussion and future directions are presented in Section IV. The conclusions are given in Section V.

II. Speech-Signal-Based Frequency Warping

The goal of this letter is to obtain a frequency warping based purely on the properties of the acoustic speech signal. This should tell us how the production system encodes the frequency information in the speech signal. It will also allow us to compare how well the speech production and auditory systems are matched. To achieve the above, utterances from a large speech corpus are analysed framewise using the short-time Fourier analysis. A periodogram estimate is computed for each frame of each utterance. An ensemble average is taken across the entire speech corpus. A logarithm of the resulting average power spectrum is divided into equal area intervals. Center frequencies of each equal area interval form the speech-signal-based frequency warping.

Suppose that our aim is to find M warped frequencies. For that purpose, let us consider a discrete-time speech signal $s(n)$. Its discrete-time short-time Fourier transform is given by

$$S(n,f) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)\mathrm{e}^{-j2\pi fm/F_s}, \qquad (1)$$

where $w(n)$ is a N samples long analysis window, F_s is the sampling frequency and f denotes the continuous frequency variable. The periodogram-based power spectral estimate for the speech signal $s(n)$ is given by

$$P(n,f) = \frac{1}{N}|S(n,f)|^2. \qquad (2)$$

An ensemble average is computed by averaging $P(n,f)$ over the entire speech corpus,

$$\bar{P}(f) = \Big\langle P(n,f) \Big\rangle. \qquad (3)$$

A logarithm of the resulting ensemble spectrum is then divided into equal area intervals, such that

$$A_i = \int_{f_i}^{f_{i+1}} \log \bar{P}(f)\, df, \qquad i = 1, \ldots, M \qquad (4)$$

and

$$A_i = A_{i+1}, \qquad i = 1, \ldots, M-1 \qquad (5)$$

where A_i denotes the area of the ith interval, while f_i and f_{i+1} denote the lower and upper cutoff frequencies of the interval, respectively. Note that $f_1 = f_{\text{min}}$ and $f_{M+1} = f_{\text{max}}$, where f_{min} and f_{max} are the lower and upper frequencies of the frequency range under consideration. The M point speech-signal-based frequency warping function is given by

$$W\left(\frac{f_i + f_{i+1}}{2}\right) = \frac{i}{M}, \qquad i = 1, \ldots, M \qquad (6)$$

where $W(f)$ function becomes continuous as $M \to \infty$ and $0 \leq W(f) \leq 1$.

III. Experiments and Results

A. Speech Corpora

In our investigations, we employ two popular speech corpora, namely TIMIT and resource management (RM). Both databases consist of speech sampled at 16 kHz. The TIMIT speech corpus is composed of 6300 utterances spoken by 630 speakers [16], while the RM corpus consists of 21000 utterances spoken by 160 speakers [17].

B. Speech-Signal-Based Frequency Warping

To obtain the speech-signal-based frequency warping we first compute an ensemble average short-time power spectrum, $\bar{P}(f)$, over each corpus outlined in Section III-A. To achieve this, the procedure detailed in Section II is employed. The frame duration is set to 25 ms and the frame shift is set to 10 ms. The Hamming window is used as the analysis window. The FFT length of 1024 samples is employed. The plots of $\log \bar{P}(f)$ as a function of frequency for each corpus are shown in Fig. 1(a). As can be seen, the two curves are very similar even though the TIMIT and RM corpora are composed of speech belonging to different speakers.

The $\log \bar{P}(f)$ curves are then divided into equal area intervals by computing their corresponding cumulative log power spectra $C(f)$, defined as

$$C(f) \triangleq \int_0^f \log \bar{P}(\lambda) d\lambda. \qquad (7)$$

The resulting $C(f)$ functions are shown in Fig. 1(b). Note that equal lengths along the ordinate axis of $C(f)$ map to equal area intervals along the abscissa of $\log \bar{P}(f)$.

The speech-signal-based frequency warping, $W(f)$, is obtained by normalizing $C(f)$ as follows

$$W(f) = \frac{C(f)}{C(f_{\text{max}})}. \qquad (8)$$

Comparison of the mel and bark warping functions with the proposed speech-signal-based warpings is shown Fig. 1(c). It can be seen that the proposed warpings are similar to the auditory-based scales, indicating that the human speech production and auditory systems are well matched.

C. Automatic speech recognition experiments

The speech signal is a vehicle primarily for the transmission of linguistic speech information. It is generally assumed that the auditory and production mechanisms are optimised for this purpose; *i.e.*, these systems have evolved over time to maximise the transmission of linguistic information necessary for correct speech recognition. If this premise is correct, then cepstral features incorporating the proposed speech-signal-based frequency warping should work well for ASR.

In this section, we compare the performance of the traditional mel-frequency cepstral coefficients (MFCCs) with cepstral features based on the proposed frequency warping in a speaker-independent ASR task. We refer to cepstral features obtained using the speech-signal-based frequency warping as speech-signal-based frequency cepstral coefficients, or SFCCs.

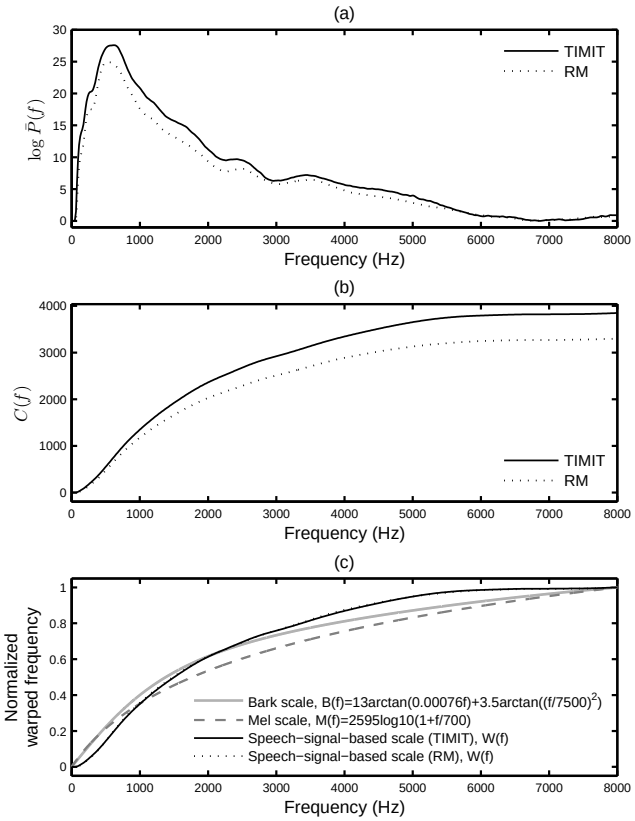

Fig. 1. Experimental results: (a) logarithm of the average short-time power spectrum, $\log \bar{P}(f)$, computed over the TIMIT corpus (black solid line), as well as the RM corpus (black dotted line); (b) the cumulative function $C(f)$ for the TIMIT corpus (black solid line), as well as the RM corpus (black dotted line); (c) normalized frequency warping functions for the bark scale (grey solid line), the mel scale (grey broken line), as well as for the proposed speech-signal-based warpings computed from the TIMIT corpus (black solid line) and the RM corpus (black dotted line).

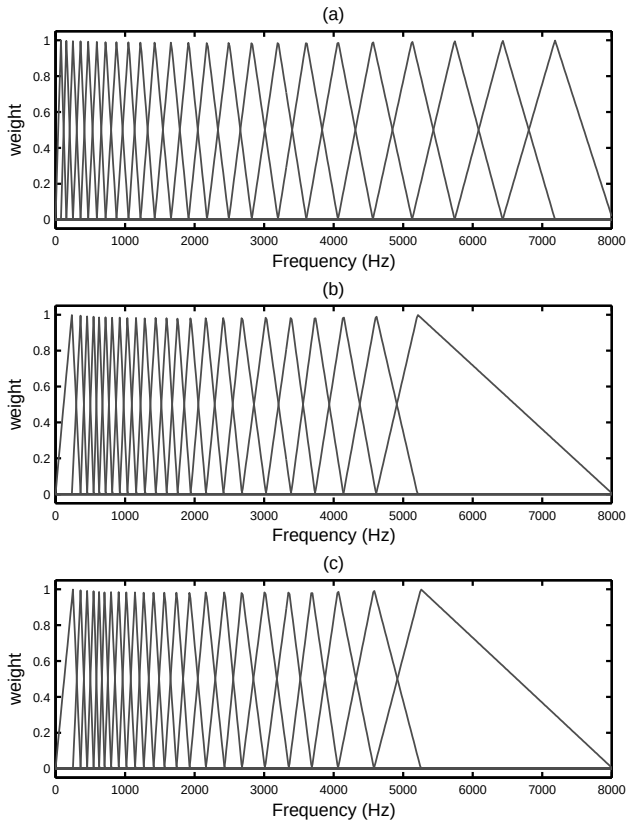

Fig. 2. Triangular filterbanks, used for cepstral features computation, with uniform filter spacing on (a) the mel scale; as well as the speech-signal-based scales obtained from (b) the TIMIT corpus and (c) the RM corpus.

The SFCCs are computed using the MFCC procedure [9]; however, the speech-signal-based frequency warping is used for the triangular filterbank design instead of the mel scale. A triangular filterbank with uniformly spaced filters on the mel scale is shown in Fig. 2(a), while filterbanks designed using speech-signal-based frequency warping, from the TIMIT corpus and the RM corpus, are shown in Fig. 2(b) and Fig. 2(c), respectively. It can be seen that the speech-signal-based filterbanks are almost identical. This is due to the similarity of the TIMIT and RM based warping functions. It can also be seen that the speech-signal-based filterbanks are somewhat similar to the mel filterbank, with notable differences being that in the speech-signal-based filterbanks denser sampling occurs at the lower frequencies, while the high frequency content is sampled more sparsely.

The ASR experiments were conducted on the TIMIT speech corpus using a setup similar to the one given in [18]. The results of the ASR experiments, in terms of phoneme recognition accuracy (%) [19], are shown in Table I. The SFCCs achieve ASR performance comparable to that of MFCCs under clean conditions. While the MFCCs are based on the mel scale (which has been obtained through diligent perception experiments) the SFCCs achieve comparable performance by using the speech-signal-based frequency warping. In addition SFCCs offer robustness improvements in noise over MFCCs. This can be attributed to how the energy is distributed across the speech spectrum. The higher frequencies of the speech spectrum have less energy and are thus more susceptible to noise. Averaging over wider frequency bands gives more reliable estimates at higher frequencies. Thus, in the proposed approach the noise does not affect high frequency components as much as it does in the mel warping case, since in the proposed approach it gets sampled more sparsely.

IV. Discussion

The approach proposed in this study can be viewed as a sampling operation. The sampling should be done judiciously; *i.e.*, equal importance should be given to equal energy regions. In the proposed approach we determine equal importance regions based solely on the basis of the log speech power spectrum. We do so in such a way that the regions with more power are sampled more finely, while the regions with less power are sampled more sparsely. That is, we sample the frequency axis non-uniformly with high energy regions sampled more densely than the low energy regions. Since the high energy spectral regions have higher SNRs than the low energy regions, our approach provides a robust way to compute the frequency warping function.

TABLE I
TIMIT ASR RESULTS IN TERMS OF PHONEME RECOGNITION ACCURACY (%) FOR WHITE NOISE

FEATURES	FREQUENCY WARPING	SNR (dB)	ACCURACY (%)								
			0	**5**	**10**	**15**	**20**	**25**	**30**	**35**	∞
MFCC	mel scale		7.73	14.76	22.75	31.36	42.98	54.68	63.24	68.73	70.59
SFCC	speech-signal-based scale (RM)		9.60	16.60	25.60	36.44	48.22	58.45	64.81	68.67	70.21
SFCC	speech-signal-based scale (TIMIT)		10.41	17.11	25.24	36.77	48.84	58.50	64.87	68.95	70.28

Our speech-signal-based (or data-driven) approach results in a warping function which is similar to the mel and bark scales, as shown in Fig. 1(c). Note that the data-driven approach proposed by Umesh *et al.* [13]–[15] also produces a mel-like warping function. However, their approach differs from ours in terms of the criterion used for computation of the warping function. Umesh *et al.* obtained the warping function such that the warped spectrum of two enunciations of the same speech sound from two different speakers are shifted versions of one another. Our approach computes the warping function such that the higher energy spectral regions are sampled more densely than the lower energy regions.

Note that in this work, we are using energy spectrum (*i.e.*, an energy vs. frequency function) to derive the warping function. This is appropriate when the background noise is white in nature. If it is coloured, then the warping function has to be computed using an SNR spectrum (*i.e.*, an SNR vs. frequency function). For this, we need to know the noise spectrum, which can be computed using a noise estimation algorithm [20].

The proposed approach provides an objective and simple way to determine a frequency warping function directly from the speech signal. It can be easily tailored for individual speakers, genders, languages, etc., as long as these are known a priori. A frequency warping function tailored for individual speakers, for example, could then be employed for construction of speaker-dependent filterbanks, which have good application potential in two areas: speaker verification and speaker-dependent ASR. Note that the proposed approach could also be employed to derive frequency warping tailored for individual frames under framewise speech processing.

V. CONCLUSION

In this letter, we derive a frequency warping directly from the acoustic speech signal. For this purpose, an ensemble average short-time power spectrum is computed over a large speech corpus. The frequency warping is obtained by taking a logarithm of the ensemble power spectrum and by considering equal area intervals. We show that the derived speech-signal-based frequency warping is very similar to the perception based warpings, such as mel and bark scales. The derived frequency warping is employed in filterbank design. The produced filterbank is shown to be quite similar to the standard mel filterbank. Results of our ASR experiments show that cepstral features based on the proposed warping achieve performance under clean conditions comparable to that of mel-frequency cepstral coefficients, while producing higher results under noisy conditions.

REFERENCES

[1] M. Hunt, "Spectral signal processing for ASR," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU)*, Dec 1999, pp. 17–25.

[2] H. Hermansky and J. Pavel, "Psychophysics of speech engineering systems," in *Proc. Int. Congr. Phonetic Sci. (ICPhS)*, vol. 3, Stockholm, Sweden, Aug 1995, pp. 42–49.

[3] S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," in *Proc. European Conf. Speech Commun. and Technology (EUROSPEECH)*, Rhodes, Greece, Sep 1997, pp. 409–412.

[4] H. Hermansky, "Should recognizers have ears?" *Speech Communication*, vol. 25, no. 1-3, pp. 3–27, Aug 1998.

[5] N. Malayath and H. Hermansky, "Data-driven spectral basis functions for automatic speech recognition," *Speech Communication*, vol. 40, no. 4, pp. 449–466, Jun 2003.

[6] F. Valente and H. Hermansky, "Discriminant linear processing of time-frequency plane," in *Proc. Int. Conf. Spoken Language Process. (ICSLP)*, Pittsburgh, PA, USA, Sep 2006, pp. 349–352.

[7] S. Stevens, J. Volkmann, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, Jan 1937.

[8] B. Moore, Ed., *Hearing*, 2nd ed., ser. Handbook of perception and cognition. San Diego: Academic Press, 1995.

[9] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall, 2001.

[10] L. Burget and H. Hermansky, "Data driven design of filter bank for speech recognition," in *Proc. Int. Conf. Text, Speech, and Dialogue (TSD)*. London, UK: Springer-Verlag, 2001, pp. 299–304.

[11] A. Biem and S. Katagiri, "Filter bank design based on discriminative feature extraction," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, Adelaide, Australia, Apr 1994, pp. 485–488.

[12] A. Biem, S. Katagiri, E. McDermott, and B.-H. Juang, "An application of discriminative feature extraction to filter-bank-based speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 96–110, Feb 2001.

[13] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency warping in speech," in *Proc. Int. Conf. Spoken Language Process. (ICSLP)*, vol. 1, Philadelphia, PA, USA, Oct 1996, pp. 414–417.

[14] S. Umesh, L. Cohen, and D. Nelson, "Frequency warping and the mel scale," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 104–107, Mar 2002.

[15] S. Kumar, S. Umesh, and R. Sinha, "Non-uniform speaker normalization using affine-transformation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, Montreal, Quebec, Canada, May 2004, pp. 121–124.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27 403–+, Feb 1993.

[17] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, New York, NY, USA, Apr 1988, pp. 651–654.

[18] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-37, no. 11, pp. 1641–1648, Nov 1989.

[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, 3.4 ed., Engineering Department, Cambridge University, 2006.

[20] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.