Presented at National Conference on Communications , NCC 2004, IISc Bangalore , Jan 30 -Feb 1, 2004

A Study of Frequency-Scale Warping for Speaker Recognition

Pradeep Kumar P and Preeti Rao
Dept of Electrical Engineering, IIT- Bombay
email: {pradeepp, prao@ee.iitb.ac.in}

Abstract

Most current speaker recognition systems use Mel frequency cepstral coefficients (MFCC) as the speaker discriminating features. MFCCs are typically obtained using a non-uniform filter bank which emphasizes the low frequency region of the speech spectrum. However some recent studies have suggested that middle and higher frequency regions of the speech spectrum carry more speaker-specific information. In this work, a general method to obtain cepstral coefficients based on different warped frequency scales is proposed. This method is applied to experimentally investigate the relative importance of specific spectral regions in speaker recognition from vowel sounds.

1. Introduction

Speaker recognition is the task of recognizing a person from his or her voice. A speaker recognition system has three basic functional modules (a) feature extraction, (b) speaker modeling and (c) pattern matching and decision-making.

Features derived during training from the speaker's speech are used to model the speaker. The most popular feature set has been the vector of Mel frequency cepstral coefficients (MFCC) traditionally used also in speech recognition. MFCCs are cepstral coefficients computed on a warped frequency scale based on known human auditory perception.

Speaker recognition systems seem to have carried over the legacy of speech recognition in terms of the choice of features. Although it is widely acknowledged that speech recognition and speaker recognition are complementary activities, practically the same features are used for both. Humans can identify a speaker even when listening to unidentifiable utterances. As a matter of fact, speech content and speaker identification are known to be processed in different areas of the human brain [6]. Speech comprehension is based in the left hemisphere of the brain and the right hemisphere is implicated in speaker identification. This suggests different mechanisms for the two functionalities.

A study by Sambur [2] to determine signal features that are most effective for speaker recognition, it was found that vowel formants (F2, F3, F4), F2 in nasals and the average pitch were the most effective features for speaker recognition.

MFCCs are typically computed by using a bank of triangular-shaped filters, with the center frequency of the filter spaced linearly for frequencies less than 1000 Hz and logarithmically above 1000 Hz. The bandwidth of each filter is determined by the center frequencies of the two adjacent filters and is dependent on the frequency range of the filter bank and number of filters chosen for design. But for the human auditory system it is estimated that the filters have a bandwidth that is related to the center frequency of the filter. Further it has been shown that there is no evidence of two regions (linear and logarithmic) in the experimentally determined Mel frequency scale [7].

Recent studies on the effectiveness of different frequency regions of the speech spectrum for speaker recognition [3], brought out the importance of frequency regions 0-500 Hz and 3500-4500 Hz. Also, higher frequencies were important for female speakers. A new filter bank front-end was proposed for speaker identification, giving more importance by way of narrower bandwidths to the frequency regions 0 to 1 kHz and 3 kHz to 4.5 kHz This

provided better performance than standard Mel scale filter bank [3].

An alternate to the filter bank method to achieve frequency-scale warping is via the bilinear transform. By suitably fixing a "warping factor" for the given sampling frequency, it is possible to obtain a desired warping function [8]. Thus the bilinear transform provides us a flexible method to achieve a range of warping functions. We use this framework to carry out an experimental study toward determining the optimal choice of frequency-scale warping for the speaker recognition task.

The present study is confined to speech vowels, which phonemes are known to contribute the most towards the speaker recognition [10]. Further, since it has been noted in recent literature [10] that the optimal filter bank for speaker discrimination may actually be phoneme-dependent, we present results for each of a set of selected vowels.

In the next section, the bilinear transform based frequency warping method is presented. Section 3 discusses implementation of warped cepstral coefficients and a measure to evaluate the effectiveness of the resulting features. Section 4 describes the experiment and the results. Section 5 gives the conclusion and future work.

2. Frequency-scale warping

A nonlinear warping of the frequency scale can be effected by bilinear transformation, given as the transfer function of a first-order all pass filter [5],

$$D(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \qquad \ldots. \qquad (1)$$

This mapping in the Z-plane maps the unit circle onto itself. From (1) we can obtain

$$\tilde{\omega} = \arg\left(D\left(e^{-j\omega}\right)\right) = \omega + 2\arctan\left(\frac{\alpha \sin(\omega)}{1 - \alpha\cos(\omega)}\right) \qquad \ldots\ldots \qquad (2)$$

The equation (2) determines the frequency-to-frequency mapping of the transformation. The warping factor α controls the shape of the warping function. In particular, for a given sampling frequency, α can be set to approximate the Mel scale or the Bark scale [8,9].

It may be noted that α can, in general, be complex-valued. The equation (1) is then rewritten as

$$D(z) = \frac{z^{-1} - \alpha}{1 - \alpha^{*} z^{-1}} \qquad (3)$$

α^{*} is the complex conjugate of α. The complex α can be represented as $|\alpha| e^{\angle\phi}$.

The phase with complex α will become

$$\tilde{\omega} = \arg\left(D\left(e^{-j\omega}\right)\right) = \omega + 2\arctan\left(\frac{|\alpha|\sin(\omega - \phi)}{1 - |\alpha|\cos(\omega - \phi)}\right) \qquad \ldots.. \qquad (4)$$

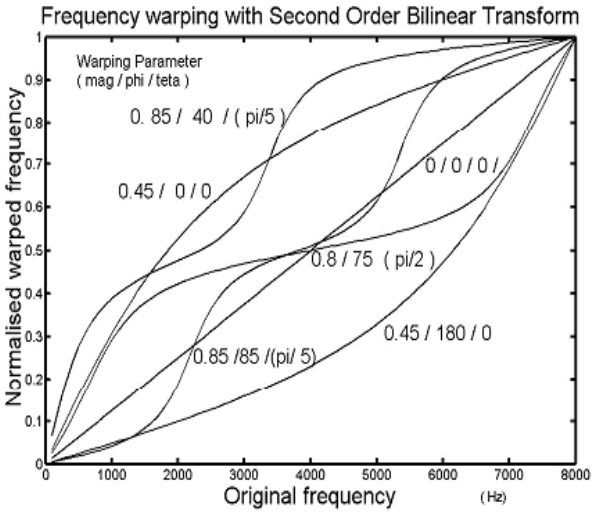

Fig.1. Frequency warping using second order Bilinear transform

More control over frequency warping can be achieved by using a second order bilinear transform [11].

$$D(z) = \left(\frac{z^{-2} - 2\alpha\cos\theta\, z^{-1} + \alpha^2}{1 - 2\alpha\cos\theta\, z^{-1} + \alpha^2 z^{-2}} \right)^{1/2} \quad .. \ (5)$$

the phase is given by

$$\tilde{\omega} = \arg\left(D\left(e^{-j\omega}\right)\right) = \omega + \arctan\left(\frac{\alpha\sin(\omega-\theta)}{1-\alpha\cos(\omega-\theta)} \right) + \arctan\left(\frac{\alpha\sin(\omega+\theta)}{1-\alpha\cos(\omega+\theta)} \right) \quad \ldots \quad (6)$$

Varying the value of θ (*between* $0 - \pi$) the warping can be controlled. With θ=0 the second order warping reduces to first order warping.

Fig.1 shows the plot of frequency warping using second order bilinear transform.. With different set of warping parameters, we can attain desired frequency warping .

Various values of α and θ are considered in the present study to gain an understanding of the importance of different frequency regions in speaker recognition.

3. Implementation

The frequency-warped cepstral coefficients of 16 kHz sampled speech signal are computed as depicted in the block diagram of Fig. 2. After pre-emphasis and windowing with a Hamming window of length 20 ms, a high-resolution DFT is used to locate the harmonic peaks and determine the spectral amplitudes at the peaks. The harmonic frequencies are frequency warped using the bilinear transformation as in equation (6). A uniform frequency-interval interpolated spectrum (of I024 points in 0-8 kHz) is calculated from the above warped envelope spectrum.

The cepstral coefficient vector is calculated from the frequency-warped spectrum using Discrete Cosine Transform

$$\tilde{c} = DCT\left(\log_{10}\left(Warped\ Amplitude\ Spectrum\right)\right)$$

The above cepstral vector is used as the feature vector (after discarding the zeroth coefficient since it represents only the signal energy).

As a measure of the effectiveness of a given feature for recognition the "divergence" is used in the study. It is a measure of distance or dissimilarity between two classes based upon information theory, and provides means of feature ranking and evaluation of class discrimination effectiveness. [1]

The divergence is defined as total average information for discriminating class ω_i from class ω_j. Assuming Gaussian distributions, the divergence between two classes is [1]

$$J_{ij} = \frac{1}{2} tr\left[\left(C_i - C_j\right)\left(C_j^{-1} - C_i^{-1}\right) \right] + \frac{1}{2} tr\left[\left(\mu_i - \mu_j\right)^T \left(C_i^{-1} + C_j^{-1}\right)\left(\mu_i - \mu_j\right) \right] \quad .. \ (7)$$

C_k , μ_k is the respective covariance matrix and means of the two classes. tr[.] is the trace of the matrix.

.

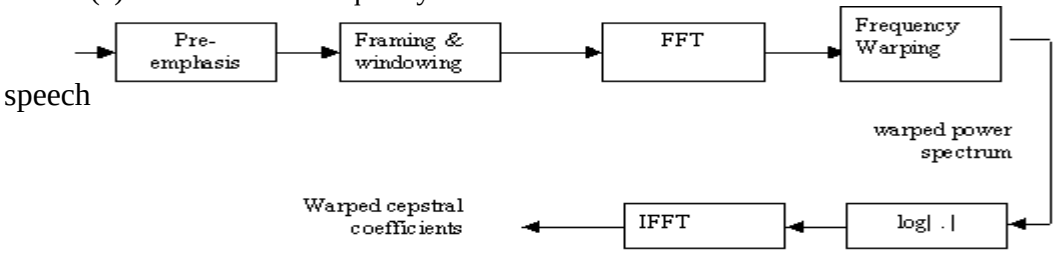

Fig 2. Computation of warped cepstral coefficients

4. Experimental results

The vowels /a/, /i/ and /oo/ were extracted from 12 words recorded from each of five different speakers at a sampling frequency of 16 kHz., with an average vowel duration of 400-500 msec

Cepstral coefficient vectors were obtained for a range of warping factors. For each warping function the effectiveness of the corresponding cepstral coefficients for feature identification was evaluated via the divergence measure applied to every pair of speakers.

At 16 kHz sampling with first order warping it was observed that the complex warping factor of 0.45 with $\varphi = 0^0$ approximates Mel scale warping and a warping factor of 0.55 approximates Bark scale warping. Second order warping was used to simulate the filter bank used by Ozgur [3] ($\theta = \pi/5 \,\text{radians}, |\alpha| = 0.85, \varphi = 40^0$). In the Ozgur warping, the frequency bands 0-1 kHz, 3 kHz – 5 kHz are enhanced and other frequency bands are compressed.

Table I shows the different warping used in the experiment.(W1-W8) Table II shows the obtained average divergence measures for the three vowels for different frequency warping. The divergence is calculated for all two-speaker combinations for different warping parameters and for 3 different cepstral dimensions.

Figure 3 shows performance of different warping function for 12 dimension cepstral feature. It is observed that the Ozgur warping [3] gives better divergence followed by Bark scale and mel scale warping. Also it is observed that the divergence measure increase linearly with Cepstral dimension for warping (W1 W4, W6 and W8)., with /i/ having higher divergence and /oo/ the least. For Bark, Mel scale and Ozgur warping (i.e W2, W3, W7) and W5 the variation is not linear. For 12 and 18 D, /i/ gives better performance.

A speaker identification test was conducted using vector quantisation as the speaker model [1]. Three utterances of each of the 3 vowels formed the training set for each speaker. The remaining 3 vowel utterances were combined to form the test sentence for each speaker. An eight code-vector VQ with various selected dimensions was used to evaluate each warping function. The average of first differences between the distortion-distance of the first and second matched speaker (in case of successful recognition) was used as a measure of discrimination between two speaker classes. (All the warping considered resulted in 100% recognition , except for one test sentence in case of 12D Mel warping).Figure 4 shows the result of this measure. Ozgur warping seems to give the best discrimination. Also it is noted that linear warping gives better discrimination than mel and Bark warping in contradiction to divergence measure. This may be due to the fact that here we are comparing only the first two matched speakers rather than averaging the divergences between speakers. Also in divergence measure we are using the entire data for each vowel, while we are clustering in VQ and only using small portion of the original data for testing.

5. Conclusion

This is an experiment conducted with speech signal sampled at 16 kHz. For this signal emphasizing the lower frequency results in improved divergence measure. Also from the better performance of Ozgur warping the importance of frequency around 3 to 5 kHz can be observed It seems that for speaker recognition there can be better warping than commonly used Mel scale warping.

This result may be valid for the individual phonemes in question, and may not hold across other phonemes. Other phonemes are to be studied, also with more speakers.

Table 1 Various Warping used in the study

Warping parameters	W1	W2	W3	W4	W5	W6	W7	W8
θ (radians)	0	0	0	0	0	$\pi/2$	$\pi/5$	$\pi/5$
$\|\alpha\|$	0	0.55	0.45	0.45	0.45	0.8	0.85	0.85
ϕ (degrees)	0	0	0	90	180	75	40	85

Table II Average Divergence Measure

		W1	W2	W3	W4	W5	W6	W7	W8
24D	/a/	355.92	20151.13	1956.41	287.70	227.27	804.76	3728881.53	231.16
	/I/	866.54	5878.15	2160.34	667.39	424.90	997.76	2049722.80	523.49
	/oo/	146.16	17070.05	2449.68	123.59	849.70	738.12	7895046.37	93.86
18D	/a/	214.88	1342.94	640.41	182.71	118.71	368.95	23693.92	165.37
	/I/	545.18	1492.86	1020.91	453.52	203.56	471.41	6312.77	397.07
	/oo/	83.19	1863.62	441.16	78.25	74.48	262.91	13216.19	60.54
12D	/a/	117.82	318.16	178.16	107.45	68.27	132.68	536.76	83.90
	/I/	244.53	550.52	523.98	219.86	122.77	195.00	665.74	232.63
	/oo/	48.91	156.21	76.91	50.97	31.96	54.18	313.02	36.07

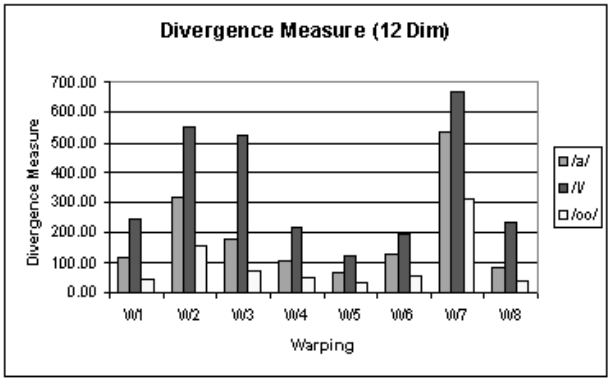

Fig 3 . Comparison of divergence

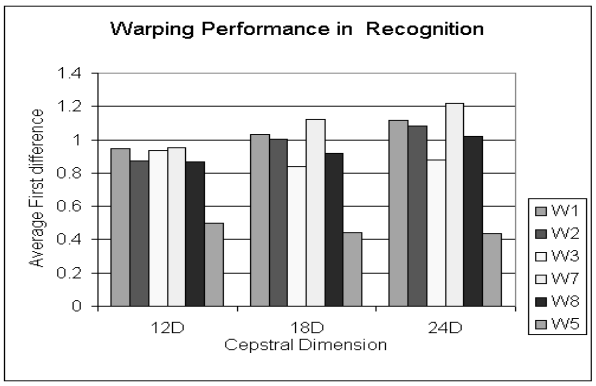

Fig 4. Recognition performance

References

[1] J P Campbell, "Speaker Recognition: A Tutorial", Proc. Of The IEEE, Vol 85, No, 9, pp. 1437- 1462, Sept 1997.

[2] M. R. Sambur, "Selection of Acoustics Features for Speaker Identification", IEEE Trans. Acoustic, Speech, Signal Processing, vol. ASSP-23, pp. 176-182, April 1975.

[3] Ozgur D Orman, " Frequency Analysis of Speaker Identification Performance", M S Thesis, Electrical and Electronics Engg., Bogazici University, Turkey, 2000.

[4] D O'Shaughnessy, " Speech Communications Human and Machine", Universities Press (India)limited, AP.

[5] A. Harma, U. K. Laine, "A Comparison of Warped and Conventional Linear Predictive Coding", IEEE Trans. On Speech and Audio Processing, vol. 9, No, 5, pp. 579 – 588, July 2001

[6] J. Wilding, S. Cook, J. Davis, "Sound familiar?" The Psychologist, Vol. 13, No.11, pp 558-62., Nov 2000.

[7] S. Umesh, L. Cohen, D. Nelson, "Fitting the Mel Scale", Proc. ICASSP, pp. 15 – 19, March 1999.

[8] J. O. Smith, J. S. Abel, "The bark Bilinear Transform", IEEE ASSP Workshop on Application of Signal Processing to Audio and Acoustics, pp 202 –205, Oct 1995

[9] K Samudravijaya, R Madan, " A novel approach to speaker verification", http://speech.tifr.res.in

[10] Eatock, J.P. and Mason, J.S.,. "A quantitative assessment of the relative speaker discrimination properties of phonemes, " Proc. ICASSP, vol I, pp. 133-136, April 1994

[11] C. Miyajima et al, "A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction", Speech Communication, vol. 35, pp 203 –218 Oct 2001