FILTER-BANK DESIGN BASED ON DEPENDENCIES BETWEEN FREQUENCY COMPONENTS AND PHONEME CHARACTERISTICS

Seyed Hamidreza Mohammadi, Hossein Sameti, Amirhossein Tavanaei, Ali Soltani-Farani

Speech Processing Laboratory, Department of Computer Engineering, Sharif University of Technology
Azadi Avenue, Tehran, Iran
{shmohammadi, tavanaei, a_soltani}@ce.sharif.edu, sameti@sharif.edu
http://spl.ce.sharif.edu/

ABSTRACT

Mel-frequency Cepstral coefficients are widely used for feature extraction in speech recognition systems. These features use Mel-scaled filters. A new filter-bank based on dependencies between frequency components and phoneme characteristics is proposed. F-ratio and mutual information are used for this purpose. A new filter-bank is designed in which frequency resolution of sub-band filters is inversely proportional to the computed dependency values. These new filter-bank is used instead of Mel-scaled filters for feature extraction. A phoneme recognition experiment on FARSDAT Persian language database showed that features extracted using the proposed filter-bank reach higher accuracy (63.92%) compared to Mel-scaled filter-bank (62.37%).

1. INTRODUCTION

Speech recognition has become an important part of many commercial systems from domestic appliance control to simple text entry systems. Efficient spectral and temporal representation of phonetic information embedded in speech waves is an important step of speech recognition systems. The Mel frequency cepstral coefficients (MFCCs) are one of the most prominent features for representing spectral characteristics of the speech signal. It is observed that higher frequency regions of the speech spectrum contain less phoneme discriminative information than low and medium frequency bands (below 3 KHz) [2]. The Mel frequency scale is an auditory scale consistent with this fact and it is used to extract the MFCCs. The frequency resolution of sub-band filters used to extract features in this manner is a decreasing function of frequency. In other words, it is assumed that a higher frequency component will always contain less information for discriminating between phonemes than a lower frequency component. Also, the frequency resolution of the Mel scale decreases in an exponential manner, which may not be true of the phoneme discriminative information embedded in speech. To solve this issue a frequency scale based on the dependencies between frequency components and phoneme characteristics is proposed. Statistical Fisher's F-ratio and Mutual Information measurements are used to measure this dependency on the FARSDAT database. Using this information, a non-uniform filter bank is designed for feature extraction.

The rest of this paper is organized in the following manner. Section 2 describes the Mel-scale and the process of computing MFCCs. Section 3 uses statistical and information theoretical methods to measure the dependencies between frequency components and phoneme characteristics. Then the proposed non-uniform filter bank is discussed. Experimental results are presented in Section 4 followed by conclusions and future work in Section 5.

2. MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Features extracted for speech recognition need to emphasize phonetic information and attenuate individual differences. Short-term Cepstral features have proven successful for this task. Cepstral coefficients may be derived from Mel-frequency log energies, or perhaps linear prediction coefficients. The former has proven more successful for robust speech recognition.

2.1 Feature Extraction

The Mel scale, proposed by Stevens, Volkman, and Newman in 1937, is the result of a numeric approximation and is based on psycho-acoustical experiments over many listeners. Simply put, the frequency components in each filter are stated to be perceived as equal frequencies by listeners. The relation between the frequency and Mel-scale is given by [5]:

$$M = 2595 log_{10}\left[\left(\frac{f}{700}\right) + 1\right] \qquad (1)$$

A Mel-scaled filter-bank with 40 filters covering the frequency range of 0 to 5512 Hz is shown in Fig. 1.

2.2 MFCC computation

For computing the MFCCs, speech signal is broken into overlapping frames and coefficients are derived for each frame of speech. The signal is pre-emphasized by applying the first order differential equation to enhance the higher frequencies of speech. In order to attenuate discontinuities at frame edges the samples in each frame are tapered by applying the Hamming windows. The length of the frames is chosen to correspond to the average duration for which the stationary assumption of speech is true; i.e. 20 to 30 milliseconds.

The Fourier transform is applied to the samples using an N-point Fast Fourier Transform (FFT). The magnitude of the spectrum is taken which is symmetric and thus half the points are sufficient. A bank of Mel-scaled band-pass filters is then used to attenuate fast changes in the spectrum and

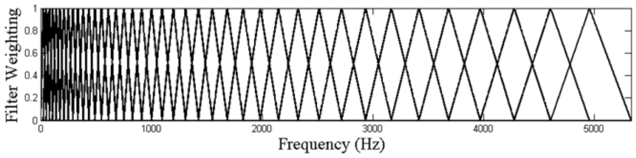

Figure 1 – A Mel-scaled filter-bank with 40 filters

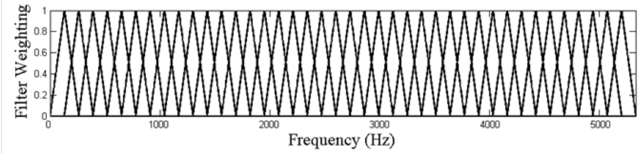

Figure 2 – A uniform-scaled filter-bank with 40 filters

estimate its envelope. Each filter is defined by its shape and frequency localization. The logarithm of the envelope multiplied by 20 gives the spectral vectors in dB.

Cepstral vectors are derived through a discrete cosine transform (DCT) of each log-spectrum vector:

$$c_j = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right) \qquad i = 1,2,\ldots,L \quad (2)$$

where is the number of log-spectrum coefficients each denoted by and is the number of Cepstral coefficients. An energy term and first and second order derivatives are usually appended to the above coefficients.

3. PHONEME DISCRIMINATIVE FREQUENCIES

To investigate the dependencies between each frequency region and phoneme characteristics, uniform-scaled filters are used to derive the energy spectrum. Each filter is a triangle-shaped band-pass filter distributed uniformly throughout the frequency spectrum. Then, the dependencies will be computed. In this paper, we describe two methods for measuring dependencies between frequency components and phoneme characteristics. The first one is mutual information which is based on calculating entropy values of certain variables. The second one is F-ratio which is a statistical method.

3.1 Uniform-scaled filters

A filter-bank with uniform-scaled filters is used for computing the power spectrum. Later, these values will be used to investigate the effect of each frequency region on discriminating phonemes.

Each filter is a triangle-shaped band-pass filter distributed uniformly throughout the whole frequency spectrum. A filter-bank with 40 uniform-scaled filters is show in Fig. 2

3.2 F-ratio for measuring the discriminative ability of frequency components

One of the methods which is usually used for finding the discriminative ability of a certain feature is F-ratio [2]. Using F-ratio, the dependency between each frequency band and phoneme statistics is computed. Each frequency band dependency will be computed independent of other frequency bands. The value computed here represents the contribution of each frequency region to phonemic discriminative information. The F-ratio is defined as [1]:

$$F = \frac{between - group\, variability}{within - group\, variability} = \frac{\frac{1}{M}\sum_{i-1}^{M}(u_i - u)^2}{\frac{1}{M.N}\sum_{i-1}^{M}\sum_{j-1}^{N}(x_i^j - u_i)^2} \quad (3)$$

where x_i^j is energy of one frequency for jth frame of phoneme with $i = 1 \ldots N$ and $j = 1 \ldots M$ and j=1…M. $u_i$and u are averages of one frequency band for ith phoneme and for all phonemes, respectively. u_i and u are defined as below:

$$u_i = \frac{1}{N}\sum_{j-1}^{N} x_i^j, \quad u = \frac{1}{M.N}\sum_{i-1}^{M}\sum_{j-1}^{N} x_i^j \quad (4)$$

F-ratio is the ratio of between-group variability (inter-phoneme variance) to within-group variability (intra-phoneme variance) in a given frequency band. So, the larger the value of the ratio for a specific frequency band, the more phoneme discriminative information that frequency band has.

3.3 Mutual Information for measuring dependencies

One of the methods for measuring the dependencies between random variables is mutual information [4]. The mutual information between two variables is defined as:

$$I(Y;P) = H(Y) - H(Y \mid P) \quad (5)$$

where Y and P are frequency band variable and phoneme class, respectively. H(.) is the entropy function, defined as:

$$H(Y) = -\sum_{y \in Y} p(y) log_2 p(y) \quad (6)$$

According to Eq. 5, $H(Y)$ is the entropy of a specific frequency band, and $H(Y|P)$ is the conditional entropy of a frequency band, given a specific phoneme. Simply put, mutual information of frequency band energy and a specific phoneme equals the reduction of the uncertainty (entropy) of that frequency band given a specific phoneme. If a frequency band has no phonemic information, $H(Y|P)$ equals $H(Y)$ and according to Eq. 5, $I(Y;P)$ will be zero, indicating that frequency band has no phonemic information. For computing $I(Y;P)$ we need to compute frequency band entropy and conditional entropy. Because these values are continuous, entropy should be computed using an estimation method. In this experiment, histogram estimation is used to calculate entropy and conditional entropy [3]. By using mutual information, we can measure the dependency between phoneme class variable and a specific frequency band variable.

3.4 Proposed filter-banks

One way to incorporate the previous section's results in designing filter-banks with more phonemic discriminative information is to assign weights to each frequency band [1]. With this method, frequency bands with higher discriminative ability will be assigned higher weights, so their impact on the final features will be more dominant. Another way is to assign higher frequency resolutions to frequency bands with higher phoneme discriminative ability. This new frequency scale can be used exactly in place of Mel-scaled filters in MFCC feature extraction process. The later approach is used in this experiment.

In real applications, the dependency measurement using F-ratio is easier than that using mutual information. In this

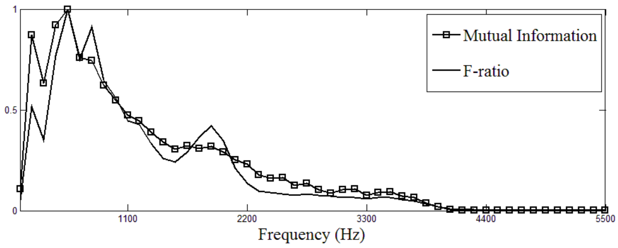

Figure 3 – The computed dependency value using F-ratio and mutual information

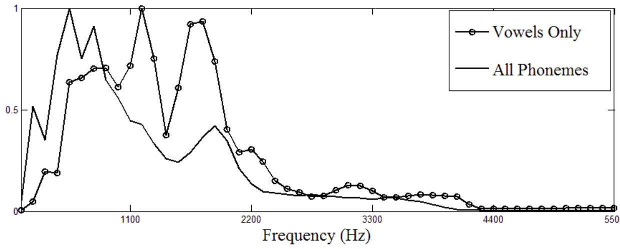

Figure 4 – The computed dependency value using F-ratio using all phonemes and only vowels

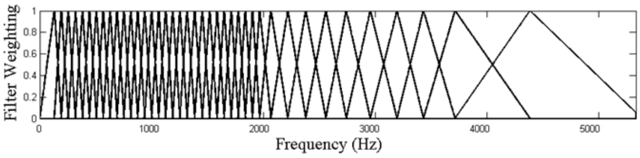

Figure 5 – The designed filter-bank based on F-ratio values

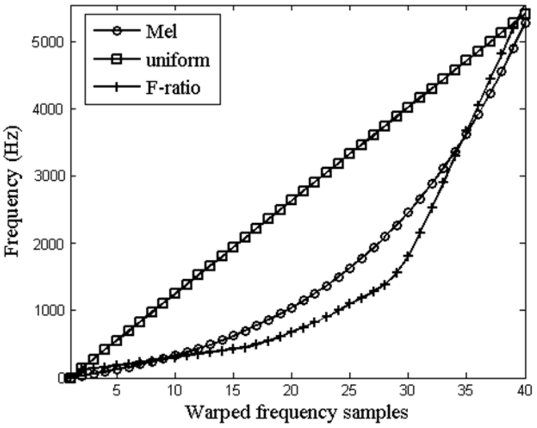

Figure 6 – Comparison of the three warping schemes

study, mutual information was also implemented to measure the dependencies but as shown in the Section 4, the results were worse than that of F-ratio. It can be considered as a measure to check the correctness of F-ratio algorithm. As it is shown in Fig. 3, both of the dependency measurement results are a little different but have a similar trend. The measured dependencies show us the fact that frequencies below 200 Hz and above 4000 Hz contain no phonemic information. This will somewhat attenuate the differences between speakers since the human pitch frequency is usually less than 300 Hz [1]. The proposed approach assigns higher frequency resolution for middle frequencies compared to Mel-scale. This is obviously because of the higher discriminative power of these frequencies, especially for vowels. One of the most important features that is relevant in discriminating between vowels are formant frequencies. The first three formants are usually in the range of 200 Hz and 3500Hz. These three formants have the most discriminative power for vowels. The higher formants are less relevant to the uttered phone and are more relevant to speaker characteristics [7]. As evident from Fig. 3, the region between 200Hz and 4000Hz has gained a reasonably high score. This is consistent with these facts, meaning that the formant frequency range is an important frequency region in discriminating between phonemes. In computing the dependency values in Fig. 3, all phonemes are considered. In Fig. 4 only 6 Persian vowels are considered for computing the dependency values using F-ratio method. The important frequencies in discriminating between vowels can be seen from Fig. 4. As can be seen, the high scores are concentrated around formant frequency regions (even more compared to dependency values for all phonemes). This is clearly because Persian vowels can easily be discriminated between by using only three formant frequencies. In Fig. 4, the trends of both curves are almost alike, but for the frequencies less than 1000Hz there exists a strong peak in the phoneme curve. This is because low frequencies have a high discriminative power for discriminating between voiced and unvoiced phonemes.

As stated earlier, each frequency band is assigned a frequency resolution according to phonetic information content by measuring the dependency between that frequency component and phoneme characteristics. Each sub-band filter's bandwidth will be assigned inversely proportional to the dependency measurement value (here, the F-ratio value). As a result, the frequency resolution of a region with high F-ratio value will be increased. The designed filter-bank is shown in Fig. 5. As one can see, it is neither uniform nor exponential in appearance. In Fig. 6, a frequency warping of uniform-scale, Mel-scale and the proposed method is plotted. It is similar to Mel scale in that both will assign high resolution to low frequencies and low resolution to high frequencies. But unlike Mel scale, the designed frequency resolution is not a strict decreasing function of frequency. As it is shown in Fig. 5, the proposed approach assigns higher frequency resolution for middle frequencies compared to Mel-scale.

Since the filter-banks are designed based on dependency values of each frequency component, it is expected that the features extracted using these new filter-banks, will improve the performance of phoneme recognition and as a consequence, improve the performance of speech recognition.

4. EXPERIMENTS AND RESULTS

For measuring the F-ratio values, FARSDAT database is used. This database is phonetically labelled. First, all samples from each phoneme are extracted from the database and a filter-bank with 50 uniform overlapping filters is applied to them. The F-ratio formula is easily computed, resulting in F-ratio values for each of the 50 filters. Mutual information is computed in the same way. A filter-bank is designed based on the computed dependency values. Then the feature vectors are computed using these newly designed filters. First, a framing of 25 ms with 10 ms shift is done. For each frame, a hamming window is applied and FFT is computed. A filter-bank consisting of 40 non-uniform sub-band filters is used to compute the power spectrum. 13 first Cepstral coefficients are computed. Energy is also added to the features. Derivation and acceleration of the features are also appended to the feature vectors. For training and testing purposes, FARSDAT (non-telephony) database is used. It contains 100 different speakers and for each speaker, about 15 minutes of read speech is recorded. In this experiment, we used all the recordings of 50 speakers. Speech from 35 speakers is used for training the phoneme models and the speech from other 15 speakers is used for testing the models. Any speakers' data which is used in training is not included in testing. FARSDAT is phonetically hand-labelled. 56 phonemes are specified in FARSDAT. In our experiments, only 30 common phonemes are considered. This is because the rest of them where mostly allophones and did not have sufficient data for training purposes.

Each phoneme is modelled by a 3-state hidden Markov model (HMM) with Gaussian mixture distributions each consisting of 16 mixtures. Phoneme models are trained for 12 iterations. The HMM toolkit (HTK) is used for this purpose [6].

A comparison is made between phoneme recognition accuracy of Cepstral coefficients extracted using uniform, Mel-scaled and the proposed filters. The result is shown in Table 1. Obviously MFCC features result in better accuracy than uniform-scaled Cepstral coefficients. As expected, features computed by using the proposed frequency scale, result in a better phoneme accuracy compared to MFCC features. Surprisingly, the frequency scale computed using Mutual Information resulted in a performance less than the F-Ratio frequency scale. This may stem from the fact that when computing Mutual Information for continuous values, an estimation step is performed, which may result in poor dependency value computation. As can be seen in Fig. 3, the Mutual Information curve is smoother compared to the F-ratio curve, which may show that Mutual Information has lower ability to detect the exact dependencies compared to F-ratio.

In another experiment, phoneme recognition accuracy was tested on only 6 Persian vowels to see the effect of this method on vowel recognition accuracy. A new filter-bank was designed based on the dependency values shown in Fig. 4 (only vowels). The vowel accuracy is shown in Table 2.

The 2.5% increase in phoneme recognition accuracy is promising since the practical upper limit for phoneme recognition accuracy is very low (about 70%).

TABLE I. PHONEME RECOGNITION ACCURACY

Frequency scale	Phoneme Accuracy
Uniform-scaled	55.20 %
Mel-scaled	62.37 %
Based on MI	63.33 %
Based on F-ratio	**63.92 %**

TABLE II. VOWEL RECOGNITION ACCURACY

Frequency scale	24 filters	40 filters
Uniform-scaled	73.67 %	71.24 %
Mel-scaled	79.34 %	84.10 %
Based on F-ratio	**83.37 %**	**89.44 %**

In addition to the above experiments, assigning weights to each Mel-scaled filter according to the F-ratio curve (instead of new frequency resolution) was also performed. These set of features resulted in accuracy almost the same as MFCCs (only a slight increase was achieved).

5. CONCLUSION

In this study, a new filter-bank was designed based on the dependencies between frequency components and phoneme characteristics. Using this new filter-bank, feature vectors were extracted and phoneme recognition accuracy was computed. The phoneme recognition accuracy showed that features extracted using the proposed filter-bank reached higher phoneme recognition accuracy (63.92%) compared to Mel-scaled filter-bank (62.37%).

REFERENCES

[1] X. Lu, J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," Speech Communication, vol.50, pp.312-322, 2008.

[2] L. Rabiner, B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall PTR, 1993.

[3] A.R. Webb, Statistical Pattern Recognition, John Wiley & Sons, 2002.

[4] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley-Interscience, 2006.

[5] J. Picone "Signal Modeling Techniques in Speech Recognition", Proc. IEEE, vol 81, no. 9, sep. 1991.

[6] S. Young, HTK Tutorial Book, http://htk.eng.cam.ac.uk/.

[7] G. Friedland, O. Vinyals, Y. Huang, C. Mueller, "Prosodic and other Long-Term Features for Speaker Diarization", IEEE Transactions on Speech and Audio Processing 17(5).985—993, 2009.