A Frequency Spectral Feature Modeling for Hidden Markov Model Based Automated Speech Recognition

Ibrahim Patel[1] and Y. Srinivas Rao[2]

[1] Assoc. Prof., Department of BME, Padmasri.Dr.B.V.Raju Institute of Technology, Narsapur Medak (Dist), A.P.
[2] Assoc. Prof., Department of Instrument Technology, Andhra University, Vizag, A.P.
Ptlibrahim@gmail.com, srinniwasarau@gmail.com

Abstract. This paper presents an approach to the recognition of speech signal using frequency spectral information with Mel frequency for the improvement of speech feature representation in a HMM based recognition approach. A frequency spectral information is incorporated to the conventional Mel spectrum base speech recognition approach. The Mel frequency approach exploits the frequency observation for speech signal in a given resolution which results in resolution feature overlapping resulting in recognition limit. Resolution decomposition with separating frequency is mapping approach for a HMM based speech recognition system. The Simulation results show a improvement in the quality metrics of speech recognition with respect to computational time, learning accuracy for a speech recognition system.

Keywords: speech-recognition, Mel-frequencies, DCT, frequency decomposition, Mapping Approach, HMM.

1 Introduction

Speech recognition is a process used to recognize speech uttered by a speaker and has been in the field of research for more than five decades since 1950s [1]. Voice communication is the most effective mode of communication used by humans. Speech recognition is an important and emerging technology with great potential. The significance of speech recognition lies in its simplicity. This simplicity together with the ease of operating a device using speech has lots of advantages. It can be used in many applications like, security devices, household appliances, cellular phones, ATM machines and computers.

With the advancement of automated system the complexity for integration & recognition problem is increasing. The problem is found more complex when processing on randomly varying analog signals such as speech signals. Although various methods are proposed for efficient extraction of speech parameter for recognition, the MFCC method with advanced recognition method such as HMM is more dominant used. This system found to be more accurate under low varying environment but fails to recognition speech under highly varying environment. This needs to the development of a efficient recognition system which can provide is efficient varying system.

N. Meghanathan et al. (Eds.): NeCoM, WiMoN, and WeST 2010, CCIS 90, pp. 134–143, 2010.

Research and development on speaker recognition method and technique has been undertaken for well over four decade and it continues to be an active area. Approaches have spanned from human auditory [2] and spectrogram comparisons [2], to simple template matching, to dynamic time-warping approaches, to more modern statistical pattern recognition [3], such as neural networks and Hidden Markov Model (HMM's) [4].

It is observed that, to extract and recognize different information from a speech signal at variable environment, many algorithms for efficient speech recognition is proposed in past. Masakiyo Fujimoto and Yasuo Ariki in their paper "Robust Speech Recognition in Additive and channel noise environments using GMM and EM Algorithm" [5] evaluate the speech recognition in real driving car environments by using a GMM based speech estimation method [6] and an EM algorithm based channel noise estimation method.

A Gaussian mixture model (GMM) based speech estimation method proposed in J.C.Segura et al [6] estimates the expectation of the mismatch factor between clean speech and noisy speech at each frame by using GMM of clean speech and mean vector of noise. This approach shows a significant improvement in recognition accuracy. However, the Segura's method considered only the additive noise environments and it did not consider about the channel noise problem such as an acoustic transfer function, a microphone characteristic etc.

A Parallel model combination (PMC) method [7] has been proposed by M.J.F Gales, and S.J.Young adapts the speech recognition system to any kinds of noises. However, PMC has a problem, of taking huge quantity of computation to recognize the speech signal. Another method for speech recognition called "spectral subtraction" (SS) is also proposed as a conventional noise reduction method [3]. However, using spectral subtraction method degrades the recognition rate due to spectral distortion caused by over or under subtraction. Additionally, spectral subtraction method does not consider the time varying property of noise spectra, because it estimates the noise spectra as mean spectra within the time section assumed to be noise.

Hidden Markov Model (HMM) [4] is a natural and highly robust statistical methodology for automatic speech recognition. It was tested and proved considerably in a wide range of applications. The model parameters of the HMM are essence in describing the behavior of the utterance of the speech segments. Many successful heuristic algorithms are developed to optimize the model parameters in order to best describe the trained observation sequences. The objective of this paper is to develop an efficient speech recognition algorithm with the existing system following HMM algorithm. The paper integrates the frequency isolation concept celled as sub band decomposition to the existing MFCC approach for extraction of speech feature. The additional feature concept of provides the information of varying speech coefficient at multiple band level this feature could enhancement the recognition approach then the existing one.

2 Hidden Markov Modeling

A Hidden Markov Model is a statistical model for an ordered sequence of variables, which can be well characterized as a parametric random process. It is assumed that

the speech signal can be well characterized as a parametric random process and the parameters of the stochastic process can be determined in a precise, well-defined manner. Therefore, signal characteristics of a word will change to another basic speech unit as time increase, and it indicates a transition to another state with certain transition probability as defined by HMM. This observed sequence of observation vectors O can be denoted by

$$O = (o(1), o(2), ...o(T))$$

Where each observation of ('t') is an m-dimensional vector, extracted at time't' with

$$O(t)=[0_1(t),0_2(t),.....,o_m(t)]^T$$

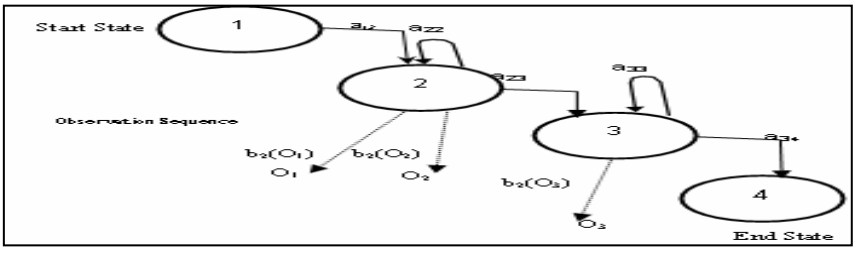


Fig. 1. A typical left-right HMM (a_{ij} is the station transition probability from state i to state j)

Figure.9 A typical left-right HMM (a_{ij} is the station transition probability from state i to state j; O_t is the observation vector at time t and bi(O_t) is the probability that O_t is generated by state i).

The HMMs is used to solve three main problems. These problems are described as following:

1: Given the model $\lambda = \{A,B, \prod\}$ and the observation sequence, how to efficiently compute $P(O|\lambda)$, the probability of occurrence of the observation sequence in the given model.

Problem 2: Given the model $\lambda=\{A,B, \prod\}$ and the observation sequence, how to choose a optimal corresponding state sequence.

Problem 3: How to adjust the model parameters $\lambda= \{A,B,\prod\}$ so that $P(O|\lambda)$ is maximized.

Problem 1 and Problem 2 are analysis problems while problem 3 is a synthesis or model-training problem. To solve these three problems, some basic assumptions are being made in HMM.

a. The output independence assumption: The observation vectors are conditionally independent of the previously observed vectors.

b. The stationary assumption: It is assumed that state transition probabilities are independent of the actual time at which the transition takes place. It can be formulated mathematically as,

$$P[q_{t1+1}= j \mid q_{t1} = i] = P[q_{t2+1}= j \mid q_{t2} = i] \quad \text{for any } t_1 \text{ and } t_2.$$

The determination of the optimal set ω of parameters in correspondence to a given utterance can be undertaken by relying on a simple property of the quantities to be maximized in both the two cases (MLE, MAP). Both the quantity to be maximized and the parameters we are looking for are probabilities, i.e. nonnegative quantities is smaller than 1. Their variations during the optimization process from the starting values to the final optimized ones are very small. As a consequence, all these variations can be considered as differentials. If Q is the quantity to be maximized and its starting and final value, after maximization, are respectively Q_{start} and $Q_{opt,}$ we can write

$$Q_{opt} - Q_{start} = Dq$$

Similarly, the variations of the parameters of the model, from the starting values to the final optimized ones, can be considered as differentials: d, $\pi_{i,}$ $da_{ij,}$ $db_i(Y_t)$, i = 1,…,N, J=1,…N, t =1,…, T.

q being a parameter, q' denoting its optimal value and q_{start} the initial value from which we start the maximization. Consequently, the determination of the optimal values of e can be simply undertaken by maximizing above equation with respect to ω' and therefore neglecting in above equation the initial values ω_{start}. The coefficients multiplying logarithms of the parameters are determined on the basis of Y_T and ω_{start}. The maximization procedure initially requires modeling densities $b_i(y)$. Among the several possible solutions, the most used is based on mixtures of Gaussian functions and the $b_i(y)$ is themselves constrained to,

$$\int b_i(Y)\, dy = 1 \; ; \; \int b_{ik}(Y)\, dy = 1$$

The above model is reasonable on the basis of the regularization theory applied to the approximation of unknown mappings, as is the case in the present situation. The consequence of this model on function a (ω, ω') is that of modifying the term where the output probabilities appear.

3 Mel Spectrum Approach

Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency 'f' in Hz;

$$\text{mel}(f) = 2595 * \log_{10}(1 + f / 700)$$

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel scale where the filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The modified spectrum of S(ω) thus consists of the output power of these filters when S(ω) is the input. The number of Mel spectrum coefficients, K, is typically chosen as 20.

Note that this filter bank is applied in the frequency domain; therefore it simply amounts to taking those triangle-shape windows in the Figure 1 on the spectrum. A useful way of thinking about this Mel-wrapping filter bank is to view each filter as an histogram bin (where bins have overlap) in the frequency domain.

The log Mel spectrum is converted back to time. The result is called the Mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the $\tilde{S}_k, k = 1,2,...,K$ time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those Mel power spectrum coefficients that are the result of the last step are we can calculate the MFCC's, $\tilde{c}_n$, as the first component,

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], n = 1,2,3...K$$

$\tilde{c}_0$, from the DCT since it represents the mean value of the input signal, which carried little speaker specific information. As shown in the figure 3.

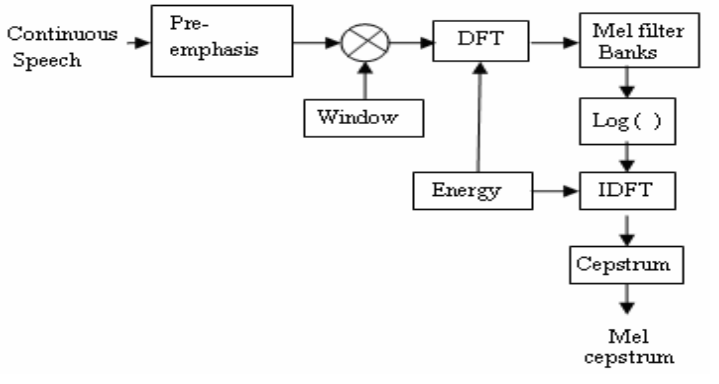


Fig. 2. Speech process models

By applying the procedure described above, for each speech frame of around 30msec with overlap, a set of mel-frequency cepstrum coefficients is computed. These are result of a cosine transform of the logarithm of the short-term power spectrum expressed on a mel-frequency scale. This set of coefficients is called an acoustic vector. Therefore each input utterance is transformed into a sequence of acoustic vectors. As these vectors are evaluated using a distinct filter spectrum the feature information obtained is limited to certain frequency resolution information only and needs to be improved. In the following section a frequency decomposition method incorporated with existing architecture is suggested for HMM training and recognition. These mel spectrum is used a recognition information in conventional speech recognition system. The spectrum doesn't exploit the variations in fundamental resolution & hence is lower in accuracy to improve the accuracy of operation a spectral decomposition approach is respected.

4 Spectral Decomposition Approach

Filter bank can be regarded as wavelet transform in multi resolution band. Wavelet transform of a signal is passing the signal through this filter bank. The outputs of the different filter stages are the wavelet and scaling function transform coefficients. Analyzing a signal by passing it through a filter bank is not a new idea and has been around for many years under the name sub band coding. It is used for instance in computer vision applications.

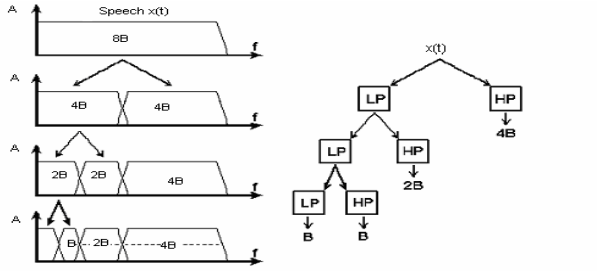


Fig. 3. Splitting the signal spectrum with an iterated filter bank

The filter bank needed in sub band coding can be built in several ways. One way is to build many band pass filters to split the spectrum into frequency bands. The advantage is that the width of every band can be chosen freely, in such a way that the spectrum of the signal to analyze is covered in the places of interest The disadvantage is that it is necessary to design every filter separately and this can be a time consuming process. Another way is to split the signal spectrum in two equal parts, a low pass and a high-pass part. The high-pass part contains the smallest details importance that is to be considered here. The low-pass part still contains some details and therefore it can be split again. And again, until desired number of bands are created. In this way an iterated filter bank is created.

Usually the number of bands is limited by for instance the amount of data or computation power available. The process of splitting the spectrum is graphically displayed in figure 4. The spectral decomposition obtained coefficient could be observed as,

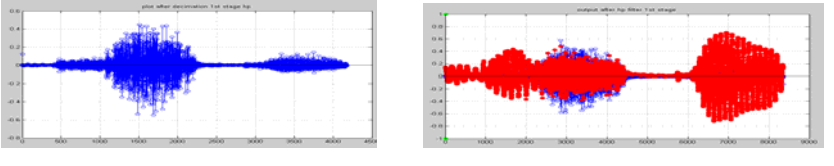


Fig. 4. Output after 1st stage decomposition for a given speech signal

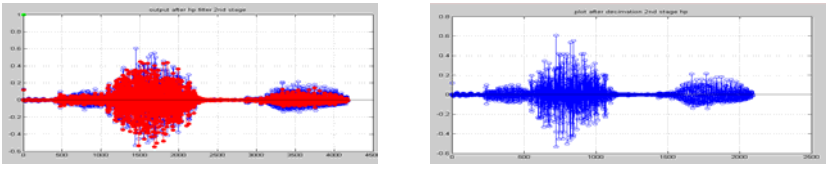


Fig. 5. Plot after 2nd stage decomposition

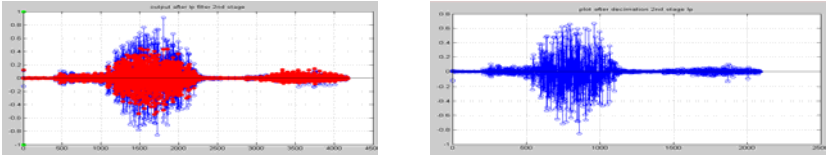


Fig. 6. Output after 3rd Stage decomposition

The advantage of this scheme is that it is necessary to design only two filters; the disadvantage is that the signal spectrum coverage is fixed.

Looking at figure 3 it is observed that it is left with lower spectrum, after the repeated spectrum splitting is a series of band-pass bands with doubling bandwidth and one low-pass band. The first split gave a high-pass band and a low-pass band; in reality the high-pass band is a band-pass band due to the limited bandwidth of the signal. In other words, the same sub band analysis can be performed by feeding the signal into a bank of band-pass filters of which each filter has a band width twice as wide as its left neighbor and a low-pass filter. The wavelets give us the band-pass bands with doubling bandwidth and the scaling function provides with the low-pass band. From this it can be concluded that a wavelet transform is the same thing as a sub band coding scheme using a constant-Q filter bank. It can be summarized, as in implementation of the wavelet transform as an iterated filter bank, it is not necessary

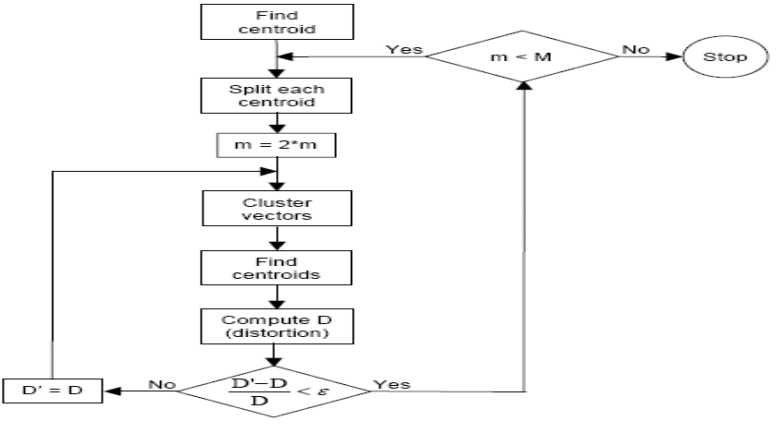
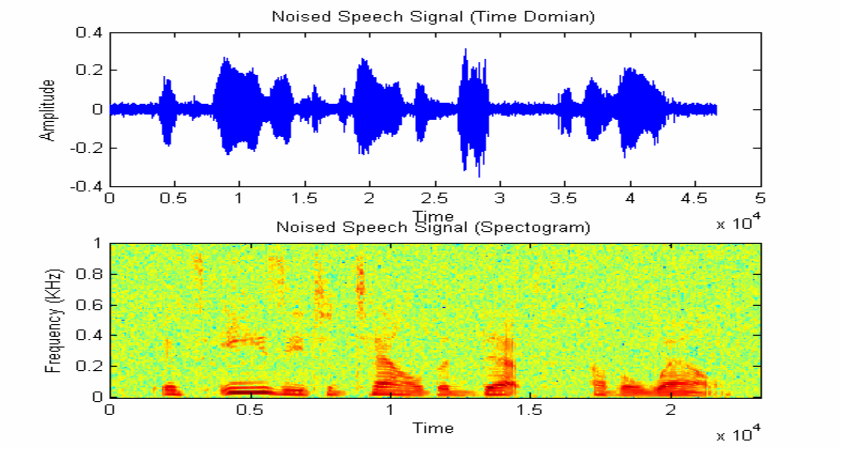
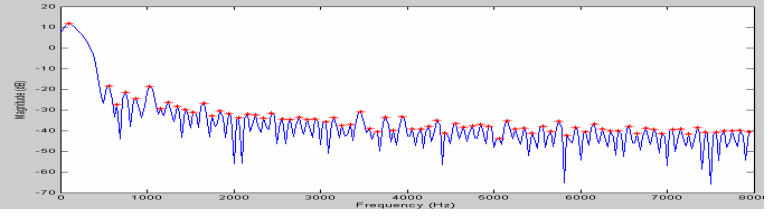


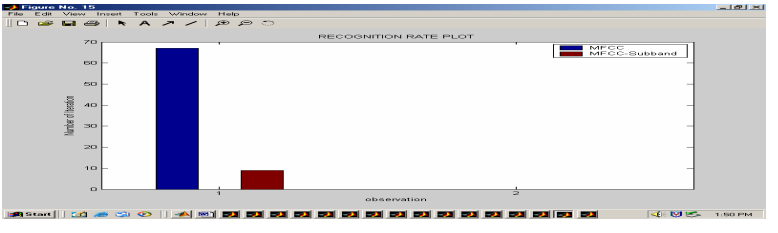
Fig. 7. Flow diagram of the LBG algorithm



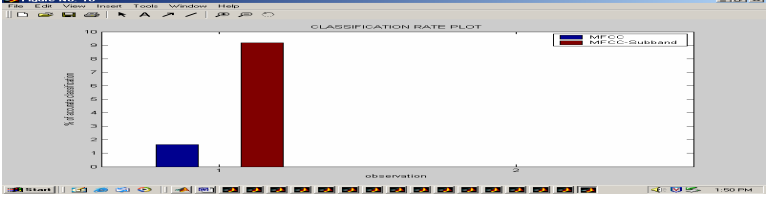
(a)



(b)

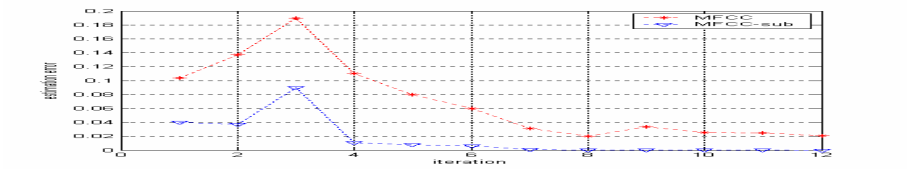


(c)

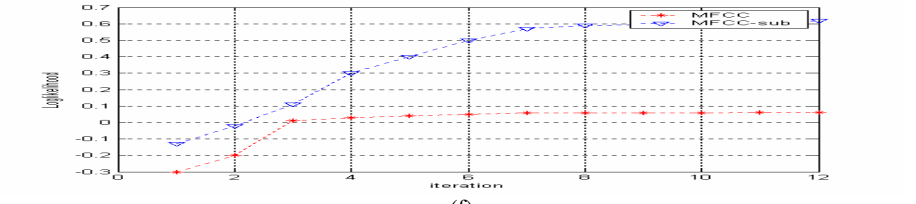


(d)

Fig. 8. (a) original speech signal and it's noise effect speech signal, (b) the energy peak points picked for training, (c) the recognition computation time for the MFCC based and the modified MFCC system, (d) the observed correct classified symbols for the two method, (e) the estimation error for the two methods wrt. Iteration, (f) the likelihood variation wrt iteration for the two methods.



(e)



(f)

Fig. 8. (*continued*)

to specify the wavelets explicitly The actual lengths of the detail and approximation coefficient vectors are slightly more than half the length of the original signal. This has to do with the filtering process, which is implemented by convolving the signal with a filter. The spectral decomposition reveals the accuracy of individual resolution which was not explored in mel spectrum. This approach is developed with a mapping concept for speech recognition as outlined between evaluation of the suggested system for a simulation of the purposed system is carried out on mat lab tool & the resolution obtained are as outlined below.

5 Mapping Approach

After the enrolment session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. A set of 'L' training vectors from the frequency information is derived using well-known LBG algorithm [4]. The algorithm is formally implemented by the following recursive procedure. The LBG algorithm designs an M-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained. Figure 7 shows, in a flow diagram, the detailed steps of the LBG algorithm. "Cluster vectors" is the nearest-neighbor search procedure, which assigns each training vector to a cluster associated with the closest codeword. "Find centroids" is the centroid update procedure. "Compute D (distortion)" sums the distances of all training vectors in the nearest-neighbor search so as to determine whether the procedure has converged. A general operational flow diagram of the suggested LBG mapping approach is shown in fig.7.

6 Simulation Observation

For the training of HMM network for the recognition of speech a vocabulary consist of collection words are maintained. The vocabulary consists of words given as,

"DISCRETE", "FOURIER", "TRANSFORM", "WISY", "EASY", "TELL", "FELL", "THE", "DEPTH", "WELL", "CELL", "FIVE", each word in the vocabulary is stored in correspondence to a feature define as a knowledge to each speech word during training of HMM network. The features are extracted on only voice sample for the corresponding word.

The test speech utterance: used for testing given as "its easy to tell the depth of a well", at 16KHz sampling.

7 Conclusion

A speech recognition system for robust to noise effect is developed. The MFCC conventional approach & extracting the feature of speech signal at lower frequency & is modified in this paper. An efficient speech recognition system with the integration of MFCC feature with frequency sub band decomposition using subband coding is proposed. The two features passed to the HMM network result in better recognition compared to existing MFCC method. From the observation made for the implemented system it is observed to have better efficiency for accurate classification & recognition compared to the existing system.

References

1. Varga, A.P., Moore, R.K.: Hidden Markov Model decomposition of speech and noise. In: Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, pp. 845–848 (1990)
2. Allen, J.B.: How do humans process and recognize speech. IEEE Trans. on Speech and Audio Processing 2(4), 567–577 (1994)
3. Kim, W., Kang, S., Ko, H.: Spectral subtraction based on phonetic dependency and masking effects. IEEE Proc.- Vision, Image and Signal Processing 147(5), 423–427 (2000)
4. Elliott, R.J., Aggoun, L., Allen, J.B.: Moore Hidden Markov Models: Estimation and Control. Springer, Heidelberg (1995)
5. Fujimoto, M., Riki, Y.A.: Robust speech recognition in additive and channel noise environments using GMM and EM algorithm. In: Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing ICASSP 2004, May 17-21, vol. 1 (2004)
6. Segura, J.C., de la Torre, A., Benitez, M.C., Peinado, A.M.: Model Based Compensation of the Additive Noise for Continuous Speech Recognition. In: Experiments Using AURORA II Database and Tasks, EuroSpeech 2001, vol. I, pp. I–941–944 (2001)
7. Gales, M.J.F., Young, S.J.: Robust Continuous Speech Recognition Using Parallel Model Combination. IEEE Trans. Speech and Audio Processing 4(5), 352–359 (1996)
8. Renals, S., Morgan, N., Bourlard, H., Cohen, M., Franco, H.: Connectionist Probability Estimators in HMM Speech Recognition. IEEE Trans. on Speech and Audio Processing 2(1), 161–174 (1994)
9. Neto, J., Martins, C., Almeida, L.: Speaker-Adaptation in a Hybrid HMM-MLP Recognizer. In: Proceedings ICASSP 1996, Atlanta, vol. 6, pp. 3383–3386 (1996)
10. Furui, S.: Digital speech processing, synthesis and recognition, 2nd edn. (2001)