Robust ASR System by Filterbank Optimization using PIO

Shubhanshi Singhal

Assistant Professor, Department of Computer Engineering, TERii, Kurukshetra, India

Vishal Passricha

Assistant Professor, Department of Computer Engineering, National Institute of Technology, Kurukshetra

ABSTRACT- Automatic Speech Recognition (ASR) system is generally divided into two parts. Front-end extracts the acoustic features from the raw speech and back-end classifies the acoustic featuresinto corresponding text. Mel-frequency cepstral coefficients (MFCCs) is widely used feature extraction technique. It derives the acoustic features by logarithmic spectral energies of the speech signal using Mel-scale filterbank. In MFCC filterbank analysis, it is observed that no consensus for spacing and number of filters is defined in various noise conditions. This paper proposes a novel approach to optimize the parameters of MFCC filterbank like central and side frequencies using pigeons inspired optimization. All the experiments are conducted on TIMIT dataset for phoneme recognition and results show that the new optimized feature set performs better than conventional MFCC.

KEYWORDS—*Automatic Speech Recognition, Filterbank, Hidden Markov Model, MFCC, Pigeons Inspired Optimization.*

1. INTRODUCTION

The primary goal of automatic speech recognition (ASR) system is to map human speech into the corresponding text without any intervention of a human. The designing of high-performance ASR system is a challenging task because there are a lot of variabilities present in speech i.e. different speaker, age, accent, background noise etc. Although various researchers claim that they have developed such systemsbut in noisy environments, the performance of their systems deteriorates drastically. This happens because these systems are highly sensitive to addictive background noise, room reverberations, and speaker variations.

For noisy conditions, various noise suppressions and spectral enhancement techniques have been proposed such as spectral subtraction [1], Kalman filtering [2], RASTA and its variants [3, 4], Weiner filtering [5], and vector Taylor series approximation [6]. These approaches are widely divided into three broad categories:

- Filtering of the noisy speech prior to classification.
- Adaptation of the speech models to include the effects of noise.
- The use of the features those are more robust to noise.

Invariant and robust features can effectively raise the robustness of ASR because they minimize the observation variability caused by the different types of inferring factor and they also reduce the possible mismatch between training and testing conditions. Mel-frequency cepstral coefficients (MFCC) features are widely used features that are generated by first applying short-time Fourier transform and then filterbank analysis on Mel-scale frequency spectrum [7]. However, the issue with these features is: first, high sensitivity toward background noise. It deteriorates the performance of MFCC-based ASR systems in noisy conditions. Second, the number of filters and their bandwidth are not standardized. 20 to 40 filters are generally used to implement MFCC that are linearly spaced before 1 KHz. In this paper,a novel approach is proposed to optimize the MFCC features. Pigeons Inspired Optimization (PIO) algorithm is a novel swarm intelligence algorithm proposed by Duan & Qiao in 2014 [8]. It can be applied for optimizing the number of filters and their spacing to improve the performance in both clean and noisy environments. The PIO optimized MFCC features are evaluated on TIMIT corpus for phoneme error rate (PER) and observed better than traditional features.

The remaining of this paper is organized as follows: section 2 covers related work i.e. optimization techniques used at front-end and back-end. Section 3 briefly discusses the MFCC, PIO, and HMM. Section 4 presents PIO-based MFCC filterbank optimization technique. Section 5 describes the experimental setup used and results obtained. Finally, this paper is concluded in section 6 with a brief discussion.

2. RELATED WORK

State-of-the-art ASR systems work as two separate modules. Front-end module extracts the useful acoustic features from the raw speech and back-end module performs

the likelihood evaluation of these features. At the back-end, GMM/HMM combination is generally used for mapping acoustic features into corresponding text. The two major issues with ASR are i) How to select feature vectors, and ii) How to set the parameters and topology. Various optimization methods have been proposed by the researchers for both ends. At earlier years, Baum-Welch and gradient methods were used to optimize the model parameters. These both models are based on hill-climbing algorithm hence strongly depend on the initial estimates of the model parameters which is a major drawback of these methods. This issue was resolved by adopting a k-means segmental procedure to derive Gaussian means and covariance parameters.

Firstly, Kwang et al. [9]introduced genetic time working in 1996 that solves non-linear time warping problem using genetic algorithm. Chau et al. [10]introduce the idea of the GA-based HMM training in 1997 which offered the better quality solutions than Baum-Welch algorithm. In this, GA optimizes HMM model parameters during HMM training. In 2001, Kwong et al. [11]extended their previous work by optimizing the HMM model parameters in a single step and found the optimal number of states for the word model. Kwang et al. [12] also proposed GA-based HMM training for MCE framework. It is found better than standard MCE methods. It also overcomes the shortcoming of traditional MCE methods that the smoothing of the empirical classification error. In 2010, Najkar et al. [13]replaced the Viterbi algorithm used in recognition phase. Data-driven design of filterbank is proposed by Burget and Hermansky [14]. Some modifications are made in standard Mel-scaled filterbank by Skowronski and Harris [15, 16] and showed improvements in recognition rate.Aggarwal and Dave [17]used combinations of feature streams and find the best filterbank by comparing different combinations. Dua et al. [18] also optimizes the filterbank by combining MFCC features with Gammatone frequency cepstral coefficients.

3. MFCC, PIO, and HMM

MFCC [7]and perceptual linear predictions[19] are popular and widely used feature extraction techniques. In both techniques, 40-dimensional feature vector is constructed by 13 coefficients + 13 first order + 13 second order time derivatives + energy. The human auditory system processes the signal in various frequency bands with linear distribution in the initial part of the frequency range and becomes non-linear towards the higher frequency range. In section 2.1, only MFCC technique is discussed in details which will be further used in experiments.

3.1 Mel-Frequency Cepstral Coefficients

MFCC is a well-proven method by researchers to extract distinct characteristics of input speech signal [20]. It uses some parts of speech production and speech perception to extract the feature vector that contains all information about the speech signal. MFCC method performs the feature extraction in the following steps:

i. Pre-emphasis and windowing: Pre-emphasis amplifies the energy of signal at high frequencies. It also reduces the differences in power components and distributes the power across the relative frequencies. Then, the input signal is partitioned into frames which contain an arbitrary number of samples. Each time frame is distributed in the overlappedhamming window to remove discontinuities from the edges. Eqn. (1) represents the mathematical formula for Hamming window:

$$W(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N} - 1\right) & 0 \le n \le N-1 \\ 0 & Otherwise \end{cases} \quad (1)$$

where $W(n)$ represents hamming window. n and N refer to current sample and total number of samples respectively.

ii. Discrete Fourier Transform (DFT): DFT is applied to divide the energy comprised into each frequency. Short-TermFourier Transform (STFT) is a fast variant of DFT and applied on each frame to extract frequency components. Eqn. (2) represents the mathematical formula for STFT:

$$f_{t,i,0} = \left|\frac{1}{N}\sum_{k=1}^{N-1}\left(e^{-j2\pi\frac{k_i}{N}}\right)f_K\right| \quad (2)$$

where $i = 0, 1, \ldots, \left(\frac{N}{2}\right) - 1$; t and N represent the time frame and a number of sampling points within a time frame t.

The spectrum obtained by STFT is filtered with low and high bandpass filter. This is required for estimating the power spectrum. Eqn. (3) represents the mathematical formula for spectrum band:

$$f_{t,k,1} = \sum_{i=0}^{\frac{N}{2}-1} c_{k,i} f_{t,i,0} \quad (3)$$

where $k = 0,1,2,\ldots, N_d$ (number of band pass filters) and c represents the amplitude of band pass filter with index k and frequency i.

iii. Cepstral Coefficients: The Mel frequency spectrum is computed using triangular shape bandpass filter. STFT obtained using eqn. (3) is used to compute the cepstral coefficients. Logarithmic Mel-Scaled filterbank is applied on Fourier transformed frame. The relation between the Mel-Scale and frequency of speech signal is given in eqn. (4).

$$Mel(f_{t,k,2}) = 2595 log_{10}\left(1 + {}^{f_{t,k,1}}/_{700}\right) \quad (4)$$

The conversion of the signal into a logarithmic form is done in order to achieve the human perception of loudness.

iv. Discrete cosine transform (DCT): DCT is applied on the Mel-coefficients to change them again into the time domain. It produces 13 MFCC features for each frame. Eqn. (5) represents the mathematical formula of DCT:

$$f_{t,k,3} = \sum_{k=1}^{N_d}\left(cos\left[\frac{i(2k-1)\pi}{2N_d}\right](f_{t,k,2})\right) \quad (5)$$

where $k = 1, 2, \ldots, N_d$; $N_c < N_d$ and N_c is the number of cepstral coefficient selected for further processing.

The speech signal is not uniform throughout the frames. To overcome this issue, cepstral coefficients over time are also added. Therefore, first (Δ) and second (ΔΔ) order derivatives are added with cepstral coefficient. Now, total 39 coefficients + energy make it complete 40-dimensional feature vector.

$$f_t = [f_{t,k,3}, \Delta f_{t,k,3}, \Delta\Delta f_{t,k,3}] + energy \qquad (6)$$

where f_t represents the final feature vector containing 39 coefficient values + energy i.e. 40 dimensional feature vector.

3.2 Pigeons Inspired Optimization

Particle Swarm Optimization, Ant Colony Optimization, Artificial Bee Colony Optimization algorithms are popular optimization algorithms. Although these optimization algorithms have remarkable performance in solving optimization problems, still there is also alarge space for improvement. In recent years, population-based swarm intelligence algorithms have been studied in depth and used in many areas to solve the optimization problem.Pigeon optimization algorithm is inspired by bio-inspired optimization based on swarm behavior like fireflies, ant, and bee which is implemented for optimization problems. In nature, pigeons find their destinations by relying on the sun, magnetic field, and landmarks. The basic PIO has two operators which are map and compress operator and landmark operator. The map and compress operator is based on magnetic field and sun, and the landmark operator is based on landmarks. The leader of the pigeon flock initiates conversation and signal to another pigeon in the flock who acknowledge back by emulating the behavior of calling pigeon and manage side by side structure emerge in a flock of definite shape. The leader of the pigeon of the flock is chosen on the basis of the number of times calls to another pigeon in the flock. A fitness function $f(x)$ attach to every pigeon that count how many times a particular pigeon called to other pigeon in given population.PIO has the capability of problem-solving. It can be used in various field of optimization like a shortest path in traveling salesman problems. PIO can also be applied to optimize the filterbank for MFCC technique.

3.3 Gaussian Mixture Model/Hidden Markov Model

GMM/HMM combination is the most successful acoustic modeling technique. Its efficient algorithm for training and recognition makes it power. GMM/HMM can efficiently model the stationary stochastic processes and the temporal relationship among the processes. This combination powers us to model dynamic speech signals using one reliable framework. Another attractive feature of GMM/HMM acoustic model is very simple to train from a given set of labeled training data (one or more sequences of observations). The two training algorithms are Baum-Welch and segmental k means both results in well-formulated and well-behaved solutions.

When the output symbols are associated with the states of the HMM then the model is known as state output HMM and when output symbols are associated with an edge then HMM model is known as edge-output HMM [21, 22]. The state output model is generally preferred over edge output model for speech recognition. A typical structure of a word based HMM is shown in figure 1. The role of acoustic modeling can be structured in a four-level hierarchy:

- Likelihood evaluation of spectral features at every HMM state.
- To find and manage the contextual phonetic variants (i.e. allophone, triphones, syllables) of the underline phoneme.
- Word composition using sub-word units (provided by HMMs) with the help of Lexicon (Pronunciation modeling).
- To generate the sequences of words or phrases up to the sentence level.

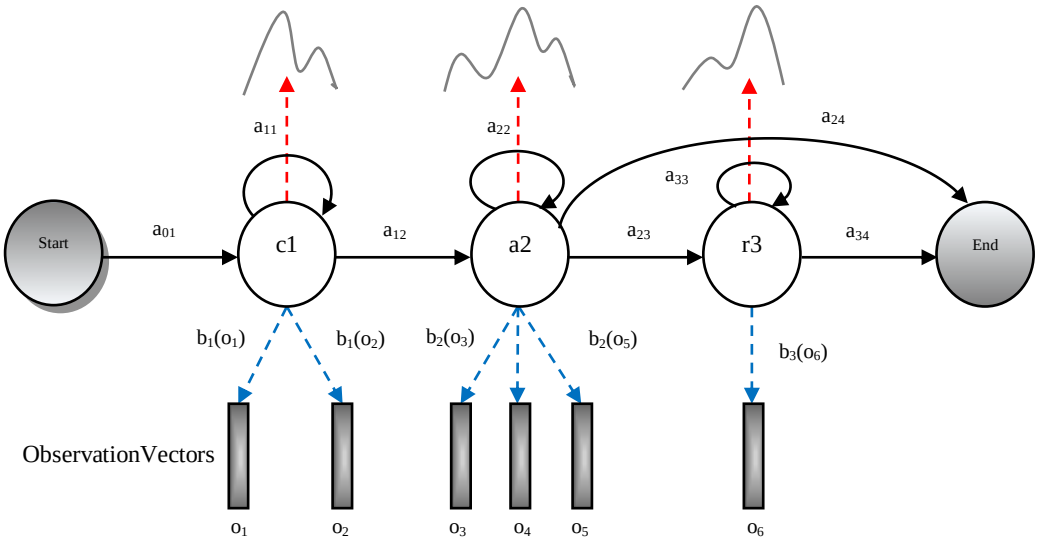

Figure 1. Block diagram of hidden Markov model

4. PROBLEM FORMULATION AND IMPLEMENTATION

In this paper, we optimized the filterbankusing PIO. Figure 2 shows the step followed in ASR for applying PIO to optimize MFCC features. The objective function used for ASR is a phoneme classifier, and the performance of this classifier is evaluated for PER on TIMIT corpus.

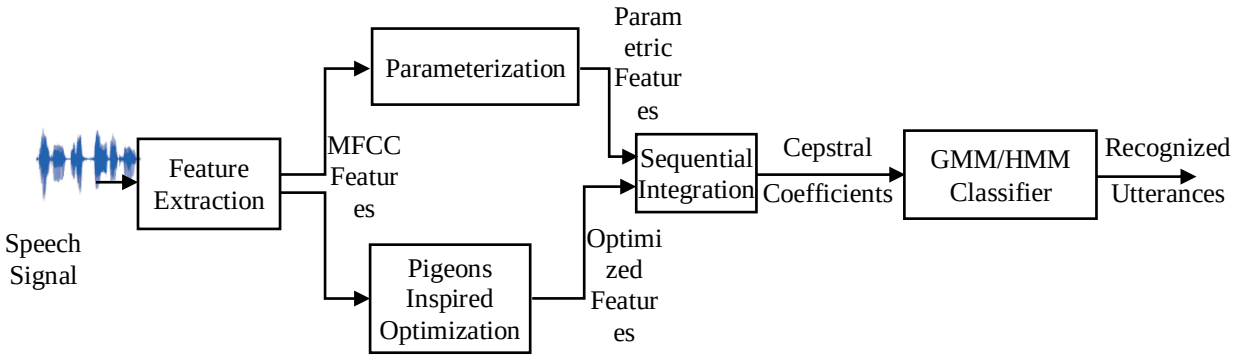

Figure 2: Block Diagram of PIO optimized MFCC based ASR System

4.1 Pigeons Inspired MFCC filterbank optimization

The mel scaled triangular filterbank is optimized by considering three parameters which correspond to the frequency values: where the triangle for the filter begins, reaches to its maximum and finally ends. Each pigeon represents a different filter and is defined as a sequence of such triangular filters represented by three frequencies as shown in figure 3.

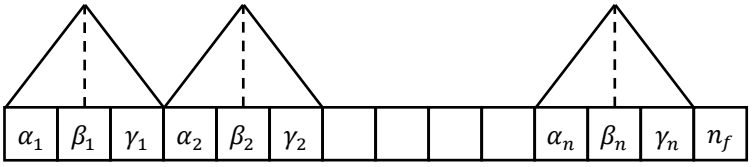

Figure 3: Frequencies for each filter in the filterbank

- Left frequency ----------α
- Center frequency--------β
- Right frequency----------γ

A filterbank is a sequence of filters and it is represented by the parameter set:

$$FB = \{F_i | i = 1, \ldots, N\} \quad (7)$$

where F_i has 3-tuple $(\alpha_i, \beta_i, \gamma_i)$ and N represents the number of filters.

Initially, a standard MFCC filterbank is generated and this filterbank is perturbed as follows:

- Randomly choose the number of filters to be modified
- Randomly choose the filters to be modified
- Change the edge frequencies of each filter selected in a small neighbor

In order to have a well-formed filterbank, the filter edge frequencies should be changed within a limit. for example the order of the frequency edges (left edge < center< right edge) should be followed. Therefore, after perturbation of a filter, this property should be checked and if it is not satisfied, perturbation should be performed again. To update a filterbank selected randomly in a small neighborhood in the parameter space, the edge of the filter (α, β, γ) can change only between -4 to +4 frequency bins.

Filterbank optimization problem has identified as natural swarm problem and can be easily implemented through the pigeons. The PIO optimizes the filterbank problem in a similar way as it solves the traveling salesman problem. A random integer $r \in [1, N]$ is produces to select the number of pigeons. Each pigeon represents a separate filter. The filter which is offering the best result is selected as leader pigeon. The leader pigeon starts conversion and signal to another pigeons about

his frequencies. Another pigeons include the leader suggestion to adjust their filters frequency. Then, they acknowledge back to their leader. The fitness function $f(x)$ is attached to every pigeon to count that how many times a particular pigeon have communicated to other pigeon in given population.

Triangular filters can be distributed along the frequency bank, with the restriction of half overlapping. This means that only the central position (parameters β_i) are required to be optimized, and the bandwidth of each filter is adjusted by the preceding and following filters.

5. RESULT AND DISCUSSIONS

5.1Experimental Setup

Human acoustic speech observations are taken from the TIMIT corpus to evaluate the performance of the PIO optimized MFCC featuresonGMM/HMM-based acoustic model. TIMIT is a standard dataset includes the utterances of both male and female speakers. It consists of 6300 utterances from the 630 speakers. We used 183 target class labels (61 phones * 3 states/phone). For decoding purpose, a phone tri-gram model is used. After decoding, the 61 phone classes are mapped into 39 useful classes as in [23]. MFCC feature extraction technique is used for extracting the features from raw speech signals. For this,the sliding window size is taken 25-ms with a fixed shift of 10-ms. 13 MFCC features + their first and second order time derivatives + energy i.e. 40 observations are supplied as input feature vector. As the number of filters $n_f$in each filterbank is not fixed, therefore, the number of output DCT coefficients is set to $\left(\frac{n_f}{2}\right) + 1$. The number of filters varies between 16 to 32. Feature extraction module, acoustic module, and decoding module havebeen developed using HTK 3.5 β-2 version toolkit. PIO is used to optimize the MFCC features. The objective of PIO is to adjust the filters frequency in maximum performance range. PIO is run for 100 iterations. An experiment is performed on a high performance computer with Intel i7-8core processor; 8GB RAM and Ubuntu 17.04 as the operating system. The noisy dataset is composed by applying a room simulator to artificially corrupt the clean signal by merging varying degrees of noise and echo so that the SNR exists between 6dB to 30dB. Restaurant lunch time recordings and YouTube are the sources of noise.

5.2 Result

At front-end, filters are derived using PIO algorithm and called as optimized filters. These filterbank are optimized in clean environment and training of system is also performed in clean environment. After optimization process, selected filterbank were tested with different levels of noise. Table 1 shows the results for PIO optimized MFCC features and normal MFCC features in PER. The results clearly indicate a significant and persistent reduction in the PER is achieved by optimizing the MFCC features. Result clearly indicates that PIO optimized MFCC features are better as compared to traditional MFCC features.PIO optimizes features performed better at lower SNRs like 6dB, 12dB, and 18dB. In clean environment, still MFCC features took lead over optimized features.

Table 1: PER for optimized MFCC features at different SNRs. Note that system is trained in clean environment

Number of filters in the filterbank	Degree of Noise (SNR)					
	6dB	12dB	18dB	24dB	30dB	Averages
OFB-32	69.25	52.55	40.49	24.12	20.13	41.308
OFB-28	69.89	52.92	41.07	24.31	22.85	42.208
OFB-24	70.02	53.85	41.96	25.08	25.36	43.254
OFB-20	70.95	54.88	42.11	32.02	27.13	45.418
OFB-16	71.48	56.19	44.16	34.53	28.93	47.058
MFCC-24	68.05	51.95	41.39	25.18	19.87	41.288

6. CONCLUSION

The main outcomes of this paper are: the various optimization techniques for front-end and back-end are discussed in section 2. This paper mainly focused on changing the filterbank bandwidths to get noise robustness. PIO technique is well-defined optimization technique, used to optimize the filterbank for MFCC features. Optimized MFCC features offered high recognition rate in both clean and noisy environments. Experimental results have illustrated that some of the filterbank performed 5% better than MFCC in a noisy environment. In future, same optimized features may be used for the large vocabulary continuous speech recognition task.

REFERENCES

1. Boll, S., *Suppression of acoustic noise in speech using spectral subtraction.* IEEE Transactions on acoustics, speech, and signal processing, 1979. **27**(2): p. 113-120.
2. Paliwal, K. and A. Basu. *A speech enhancement method based on Kalman filtering*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*. 1987. IEEE.
3. Hermansky, H. and N. Morgan, *RASTA processing of speech.* IEEE transactions on speech and audio processing, 1994. **2**(4): p. 578-589.
4. Koehler, J., et al. *Integrating RASTA-PLP into speech recognition.* in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. 1994. IEEE.
5. Chen, J., et al., *New insights into the noise reduction Wiener filter.* IEEE Transactions on audio, speech, and language processing, 2006. **14**(4): p. 1218-1234.
6. Benesty, J., M.M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. 2007: Springer.

7. Davis, S.B. and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, in *Readings in speech recognition*. 1990, Elsevier. p. 65-74.
8. Duan, H. and P. Qiao, *Pigeon-inspired optimization: a new swarm intelligence optimizer for air robot path planning.* International Journal of Intelligent Computing and Cybernetics, 2014. **7**(1): p. 24-37.
9. Kwong, S., C.-W. Chau, and W.A. Halang, *Genetic algorithm for optimizing the nonlinear time alignment of automatic speech recognition systems*. IEEE Transactions on Industrial Electronics, 1996. **43**(5): p. 559-566.
10. Chau, C.-W., et al. *Optimization of HMM by a genetic algorithm.* in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. 1997. IEEE.
11. Kwong, S., et al., *Optimisation of HMM topology and its model parameters by genetic algorithms.* Pattern recognition, 2001. **34**(2): p. 509-522.
12. Kwong, S., et al., *A genetic classification error method for speech recognition.* Signal Processing, 2002. **82**(5): p. 737-748.
13. Najkar, N., F. Razzazi, and H. Sameti, *A novel approach to HMM-based speech recognition systems using particle swarm optimization*. Mathematical and Computer Modelling, 2010. **52**(11-12): p. 1910-1920.
14. Burget, L. and H. Heřmanský. *Data driven design of filter bank for speech recognition*. in *International Conference on Text, Speech and Dialogue*. 2001. Springer.
15. Skowronski, M.D. and J.G. Harris. *Improving the filter bank of a classic speech feature extraction algorithm*. in *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on*. 2003. IEEE.
16. Skowronski, M.D. and J.G. Harris, *Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition.* The Journal of the Acoustical Society of America, 2004. **116**(3): p. 1774-1780.
17. Aggarwal, R.K. and M. Dave, *Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system.* Telecommunication Systems, 2013. **52**(3): p. 1457-1466.
18. Dua, M., R.K. Aggarwal, and M. Biswas, *GFCC based discriminatively trained noise robust continuous ASR system for Hindi language.* Journal of Ambient Intelligence and Humanized Computing, 2018: p. 1-14.
19. Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* the Journal of the Acoustical Society of America, 1990. **87**(4): p. 1738-1752.
20. Rabiner, L.R. and B.-H. Juang, *Fundamentals of speech recognition*. Vol. 14. 1993: PTR Prentice Hall Englewood Cliffs.
21. Bakis, R., *Continuous speech recognition via centisecond acoustic states.* The Journal of the Acoustical Society of America, 1976. **59**(S1): p. S97-S97.
22. He, X. and L. Deng, *A new look at discriminative training for hidden Markov models.* Pattern Recognition Letters, 2007. **28**(11): p. 1285-1294.
23. Lee, K.-F. and H.-W. Hon, *Speaker-independent phone recognition using hidden Markov models.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989. **37**(11): p. 1641-1648.