A NOVEL SPEAKER IDENTIFICATION SYSTEM

In Hwan Baek, Kayvon Sadeghi, and Mohammad Mohammad
{drfaustus, kayvon, m.mohammad}@ucla.edu

University of California, Los Angeles

ABSTRACT

We demonstrate a speaker identification system that takes input from speakers, uses feature extraction and classification, and then identifies the speaker. For the training and testing of our system, we use speech data from 10 speakers saying the vowel /a/, and the data consists of clean signals as well as noisy signals. In the system, the feature extraction consists of Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), and prosodic features. Once the features are extracted, they are then used for classification, which utilizes multiple Gaussian Mixture Models (GMMs). In order to best use the features that are extracted, we concatenate the MFCC feature matrix with the prosodic feature matrix, while keeping the LPC feature matrix separate. These matrices are then fed into separate GMMs, and a likelihood score is computed for each one. The likelihood scores from each GMM are then linearly combined in order to determine a final likelihood score. Once the final likelihood score is found, the speaker with the highest likelihood is chosen as the target speaker. Our novel system shows significant improvement over the baseline system that was provided, increasing the average result from 71% to 97.5% for a clean signal and from 65% to 90% for a noisy signal. In addition, the maximum result for a clean signal is raised from 75% to 100%, and the maximum result for a noisy signal is raised from 70% to 100%.

1. INTRODUCTION

Biometric information has been very popular for identification and authentication in security-related applications. Speaker recognition is a biometric technique that has long been investigated for such applications. It has accumulated over fifty years of progress and development and has proven to be very successful [1]. The main assumption is that the human voice is unique and can be used for recognizing speakers' identities [1].

In this report, we present a speaker identification system that employs Linear Prediction Coefficients (LPC), Mel Frequency Cepstral Coefficients (MFCC), and pitch features, which are fed into Gaussian Mixture Models for classification. The LPC and MFCC methods are used to extract information from the vocal tract. However, such methods based on the vocal tract are very sensitive to noise. The pitch features are less widely used than LPC and MFCC, but they are known to be less affected by noise. Therefore, we integrated the pitch features to improve the performance in noisy environments. We chose a Gaussian Mixture Model (GMM) as the classifier of our system. The motivation of speaker classification using GMMs is that the Gaussian components represent speaker-dependent spectral shapes, and Gaussian mixtures are capable of modeling arbitrary densities [2].

Section 2 discusses related work on speaker identification systems. We then describe the problem statement in secion 3. We discuss our system implementation in section 4. We then present our system's performance based on the data provided. Finally, we conclude by discussing our results and future work.

2. RELATED WORKS

Most speaker recognition systems consists of two main parts: feature extraction and classification.

The most popular features are ones describing the vocal tract, since it carries significant information about the user [3]. Mel Frequency Cepstral Coefficients (MFCC) are widely used to characterize the vocal tract information for speaker identity [3]. Other features that can be used to describe the vocal tract are: Real Cepstral Coefficients (RCC), Linear Prediction Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP) Coefficients, and Adaptive Component Weighting (ACW) [1]. Prosodic features that describe the glottis source can also be useful for speaker identification. For instance, the fundamental frequency and the frame energy can be utilized for speaker identification. In our system, we adopted features that describe both the vocal tract and glottis source. MFCC and LPC features were chosen for the vocal tract, while the pitch and its first and second derivatives were chosen for the glottis source.

As for the classification task, the classification approaches can be categorized into two main categories: discriminative and non-discriminative approaches [1]. Discriminative classifiers minimize the classification error and only need to model the boundary between the classes. Discriminative classifiers include Linear Discriminant Analysis (LDA), Time-Delay

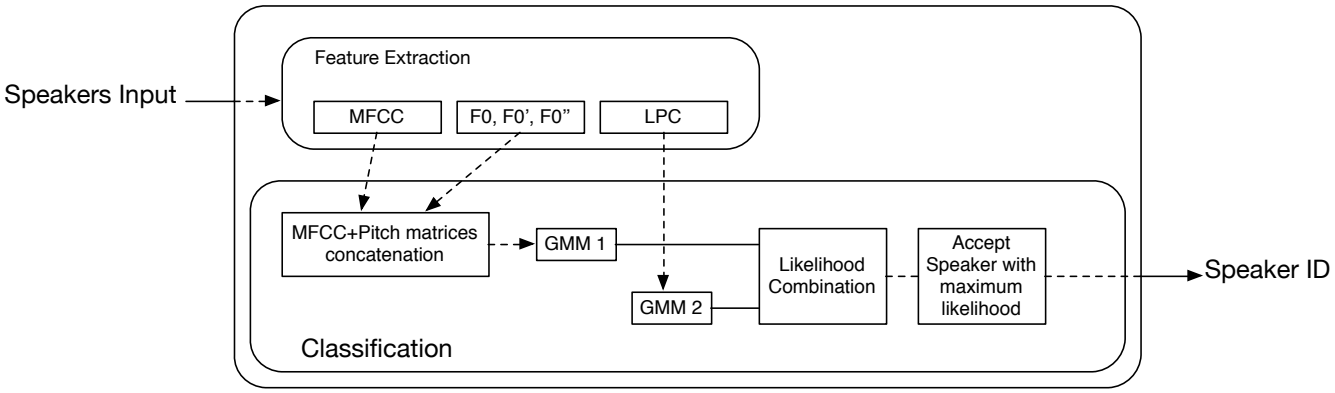

Fig. 1: A system diagram of the Speaker Identification System

Neural Networks (TDNN), and Support Vector Machines (SVM)[1]. On the other hand, non-discriminative approaches build models based on the underlying distribution of the training data [1]. Non-discriminative approaches include: Probabilistic Neural Networks (PNN), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM) [1]. For our system, we employ Gaussian Mixture models, which were already provided for us, for the classification task.

3. PROBLEM STATEMENT

We are provided with speech data that consists of 10 speakers pronouncing the vowel /a/. The speech data for each speaker has 8 utterances of the vowel. The goal is to build a speaker identification system that can identify speakers based on the data provided. The training data consists of 6 utterances per speaker and the testing data consists of 2 utterances per speaker.

Two versions of the data were provided. One version is a clean signal (i.e. speech is recorded in a noiseless environment), while the other version has 10dB noise added to it and is considered to be a noisy signal. Our system is tested with both datasets.

4. SYSTEM IMPLEMENTATION

Figure 1 shows a block diagram of our system. Training of the system is performed by first extracting the feature matrices using short-time windows with a lenght of 20ms. The features that the system utilizes for speaker identification are the following:

- Mel Frequency Cepstrum Coefficients (MFCC)
- Linear Predictive Coding (LPC) Coefficients
- The pitch (fundamental frequency), the first derivative of the fundamental frequency, and the second derivative of the fundamental frequency

Speech generation consists of two parts: the glottis source and the vocal tract. We utilize features from the two parts to enhance our system's performance. The MFCC and LPC features describe the vocal tract, while the pitch features describe the glottis source. Each feature is represented by an $m \times n$ matrix where m is determined by the number of short time windows. For the MFCC and LPC features, n is equal to p, where p is the order. For the pitch features, n is equal to 3 since we use the pitch, the first derivative of the pitch, and the second derivative of the pitch.

The MFCC and pitch features matrices are concatenated together and fed into one GMM while the LPC feature matrix is fed into another, as shown in the system diagram. The likelihood of the two GMMs are combined together and the speaker with the maximum likelihood is chosen, as will be explained in section 4.3.

4.1. Spectral Envelope features

4.1.1. Linear Prediction Coding

LPC is one of the most powerful speech analysis techniques for encoding quality speech at a low bit rate [4]. The idea behind LPC is that the current speech sample can be estimated as a linear combination of past speech samples. An all-pole filter model is used to simulate the acoustics of the vocal tract. The goal of LPC is to minimize the sum of the squared difference between the original signal and the estimated signal during a finite duration. The estimated signal is given by the equation (1):

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

where a_k is the prediction coefficient.

The implementation of our system utilizes the Matlab function lpc(x,p) where x is the input signal and p is the LPC order. To find the coefficients, this function uses the autocorrelation method, which utilizes Levinson-Durbin algorithm. Sometimes the covariance method is used for LPC instead of

the autocorrelation method, but the autocorrelation method is more widely used in practice and is faster to compute. The subsystem that processes LPC is illustrated in Figure 2.

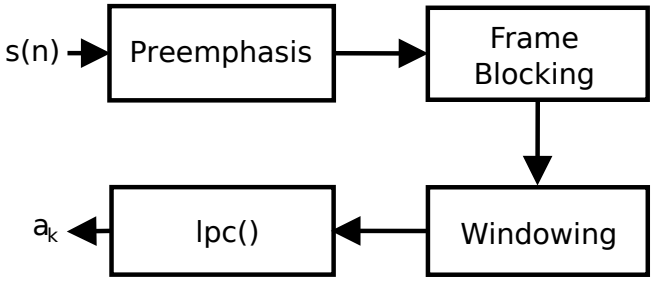

Fig. 2: The LPC subsystem

4.1.2. Mel Frequency Cepstral Coefficients

According to psychophysical studies, the human perception of sound frequencies does not follow a linear scale [2]. The human ears critical bandwidth at different frequencies has a known variation, which is the motivation behind the MFCC method [5]. Instead of using a linear scale, the Mel-frequency scale is used to capture the characteristics of speech. This is done by warping the frequency scale to mimic the frequency resolution of the auditory system spectrum. This allows for higher sensitivity to certain properties of the signal and for a resolution more similar to the resolution that a human ear would have. The Mel-frequency scale is approximately linear for frequencies below 1000Hz and logarithmic for frequencies above 1000Hz. The MFCC features correspond to the cepstrum of the energies from the different filters that are spaced uniformly on the Mel-frequency scale. [4].

We utilized the melcepst() function from Voicebox and made changes to the parameters as well as other modifications that allowed us to integrate the different functions and matrices with the other parts of our code. This new function was then used to extract the features. The process is illustrated in Figure 3.

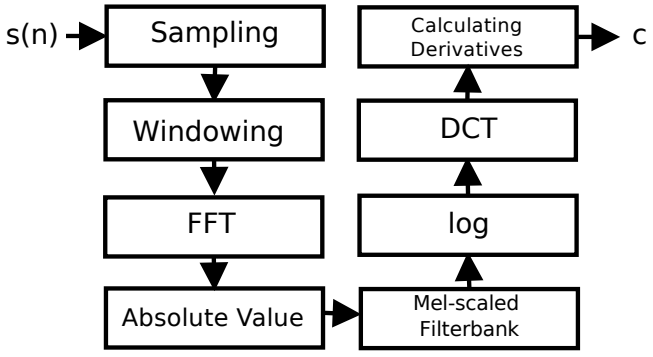

Fig. 3: MFCC

Although MFCCs are not greatly influenced by the variations of the speech waveform depending on the vocal cord condition, MFCCs are highly sensitive to noise.

For both the LPC function and the MFCC function, we use a 20ms short-time window length, 10ms window shift, and a Hamming window.

4.2. Prosodic Features

Methods based on the vocal tracts have proven to be very successful in clean conditions. However, the performance of such methods considerably deteriorates in noisy environments [3]. Therefore, various techniques have been proposed to utilize prosodic features from the glottis source as well, since they also carry information about the speaker. While they are not widely used, considering the small number of testing data, which contains only the single vowel /a/, prosodic features are good candidates for our system. In addition, prosodic features are known to be less affected by signal impairments than spectral envelope features [6]. Incorporating prosodic features in our system greatly enhanced its performance.

Reference [7] suggests using the pitch, its first and second derivative, and the first and second derivative of the power of the signal, each taken in a short-time window. We only used the first three features ($f_\circ$, as well as its first and second derivative). Reference [7] explains that the derivatives carry information about the time variations of speech that could be used for distinguishing speakers. Equation (2) is the resulting feature matrix representing the pitch parameters.

$$\text{feature} = \begin{bmatrix} f_{\circ 1} & f'_{\circ 1} & f''_{\circ 1} \\ f_{\circ 2} & f'_{\circ 2} & f''_{\circ 2} \\ \vdots & \vdots & \vdots \\ f_{\circ N} & f'_{\circ N} & f''_{\circ N} \end{bmatrix} \tag{2}$$

We adapt the algorithm described in reference [8] for pitch determination in order to design an algorithm that will work for our purposes. The algorithm estimates the pitch through spectrum shifting on a logarithmic frequency scale and calculating the Subharmonic-to-Harmonic Ratio (SHR) [8].

Theoretically, the Subharmonic-to-Harmonic Ratio is obtained as shown in equation (3):

$$SHR = \frac{\sum_{n=1}^{N} A((n-1/2)f_\circ)}{\sum_{n=1}^{N} A(nf_\circ)} \tag{3}$$

where $A(f)$ is the short term amplitude spectrum, $f_\circ$ is the fundamental frequency, and N is the maximum number of harmonics contained in the spectrum. To extract $f_\circ$, it is suggested to use an approximate computation of the SHR since equation (3) is not trivial to solve [8].

The author of the paper provided the Matlab source code, which we have utilized in the design of our system.

4.3. Combination of Methods

We investiagted the combined use of the methods in order to improve the accuracy of the system. There are two different approaches to realize this. The first approach is a straightforward combination of extracted features [9] as shown in Figure 4.

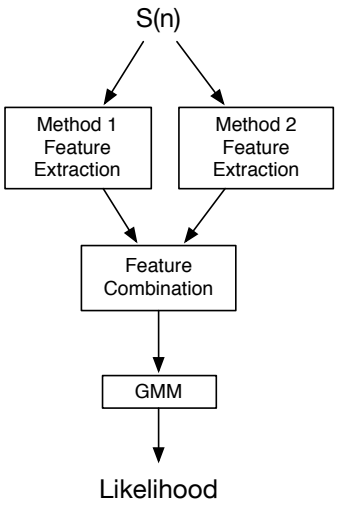

Fig. 4: Feature Combination

The second approach is the combination of the likelihood scores of the independent GMM for each method [9] as shown in Figure 5.

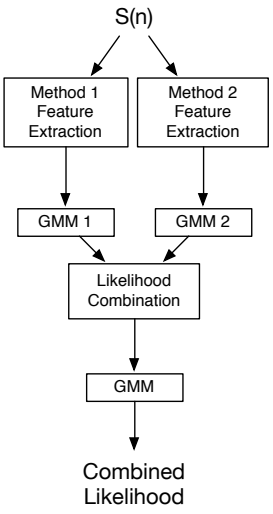

Fig. 5: Likelihood Combination

Both approaches are used to combine the LPC, MFCC, and prosodic features.

4.3.1. *Combining LPC Features with MFCC Features*

Before combining the prosodic features, the combination of just LPC and MFCC is investigated. Both the feature combination approach and the likelihood combination approach are examined. Based on our experiment results, the likelihood combination approach gives better accuracy scores. The input speech is fed into the LPC and MFCC methods separately, in order to extract features from each method. Then, the extracted features are fed into two independent GMMs. The computed likelihood scores from each GMM are combined to find a new score, as shown in equation (4):

$$L_{\text{combination}} = \alpha L_{\text{LPC}} + \beta L_{\text{MFCC}} \tag{4}$$

The likelihood of the MFCC-based GMM is linearly coupled with the likelihood of the LPC-based GMM to calculate the combined likelihood [2]. α and β are the weighting coefficients. Experiments show that choosing α = 1 and β = 1 gives a good result. After the likelihoods are combined, the speaker with the maximum likelihood is chosen as the target speaker [2].

4.3.2. *Combining with Pitch*

Integration of the prosodic features with the spectral features was a major challenge in our system. Two methods have been considered: adding a third GMM for the prosodic features, and concatenating the prosodic features matrix with either one of the spectral features matrices. The latter approach was chosen for our system over the former method. The former method resulted in poor accuracy. That is because the pitch feature matrix in (2), which contains only three columns, is not large enough for clustering.

Concatenating the pitch feature matrix with the MFCC feature matrix resulted in a higher accuracy than concatenating it with the LPC matrix. The concatenation is shown in equation (5)

$$\text{Concat}\left(\begin{bmatrix} f_{\circ 1} & f'_{\circ 1} & f''_{\circ 1} \\ f_{\circ 2} & f'_{\circ 2} & f''_{\circ 2} \\ \vdots & \vdots & \vdots \\ f_{\circ N} & f'_{\circ N} & f''_{\circ N} \end{bmatrix}, \begin{bmatrix} c_1^1 & c_1^2 & \dots & c_1^k \\ c_2^1 & c_2^2 & \dots & c_2^k \\ \vdots & \vdots & \ddots & \vdots \\ c_n^1 & c_n^2 & \dots & c_n^k \end{bmatrix}\right) = \tag{5}$$

$$\begin{bmatrix} c_1^1 & c_1^2 & \dots & c_1^k & f_{\circ 1} & f'_{\circ 1} & f''_{\circ 1} \\ c_2^1 & c_2^2 & \dots & c_2^k & f_{\circ 2} & f'_{\circ 2} & f''_{\circ 2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ c_n^1 & c_n^2 & \dots & c_n^k & f_{\circ N} & f'_{\circ N} & f''_{\circ N} \end{bmatrix}$$

5. PERFORMANCE EVALUATION

To evaluate our results, we ran each algorithm many times with the clean signal and the noisy signal. This was done so

we could determine an average and maximum value for the results. Initially, we started with the baseline system. For the clean signal, the average result was 71%, and the maximum result was 75%. For the noisy signal, the average result was 65%, and the maximum result was 70%.

When we incorporated our MFCC algorithm with the LPC algorithm, the results improved significantly. The average result for the clean signal improved from 71% to 90%, and the average result for the noisy signal increased from 65% to 78.5%. In addition, the maximum result for the clean signal jumped from 75% to 90%, while the maximum result for the noisy signal was raised from 70% to 80%.

Despite the improvement, we experimented with different approaches and modifications in order improve the result even further. The most successful modification was the incorporation of the pitch features, which showed significant improvement during the performance evaluation. The average result for the clean signal increased all the way to 97.5%, almost 30% higher than the original baseline. In addition, the average result for the noisy signal was much better, increasing to 90%, which was 25% higher than the original baseline. Using our final algorithm with the incorporation of pitch features, the maximum result for the clean signal was 100%, and the maximum result for the noisy signal was 100% as well. These results are summarized in Figure 6.

Approach	Clean	Noisy
Baseline System	AVE = 71% MAX = 75%	AVE = 65% MAX = 70%
MFCC + LPC	AVE = 90% MAX = 90%	AVE = 78.5% MAX = 85%
MFCC + LPC + pitch	AVE = 97.5% MAX = 100%	AVE = 90% MAX = 100%

Fig. 6: Results

As shown in the table, our algorithm resulted in a significant improvement over the baseline for both the clean signal and the noisy signal.

6. OTHER ATTEMPTED APPROACHES

Before we chose our final approach, we tried several other methods in order to determine the best way to design the system. These other methods are described below. First, we tried using PLP (perceptual linear prediction), which is similar to LPC, but it modifies the short-term spectrum [4]. We also tried PLP with RASTA, a modified version of PLP. That modified method aimed to smooth over short-term noise variations [4]. In addition, we modified the MFCC algorithm to create an autocorrelation mel frequency cepstral coefficient (A-MFCC) feature extraction algorithm [10]. This was done to minimize the effects of noise. Finally, we tried extraction of pitch fixtures while utilizing linear discriminant analysis (LDA) [6]. In this case, we looked at the variance, mean, skewness, and kurtosis with LDA. While some of these methods were improvements over the baseline, each of these methods had various issues, which led us to focus on other approaches.

7. CONCLUSION

In summary, we designed a novel speaker identification system that incorporates MFCC, LPC, and prosodic features in order to successfully identify the speaker. Initially, we attempted many different approaches, including PLP, PLP with RASTA, AMFCC, and pitch features with LDA, but our final system design improved over all of those methods. In our final system, after feature extraction, which includes MFCC features, LPC features, and prosodic features, our algorithm integrates the prosodic features with the MFCC features by concatenating the pitch feature matrix with the MFCC feature matrix. The LPC feature matrix is kept as a separate matrix. Then, the matrix that contains MFCC features combined with pitch features and the matrix that contains LPC features each utilize separate GMMs to determine a likelihood score. A linear combination of those likelihood scores is used to determine a final likelihood score. Then, the target speaker is chosen as the speaker with the maximum likelihood. This algorithm showed significant improvement over the baseline system, improving the average result for a clean signal from 71% to 97.5%. In addition, our algorithm improved the average result for a noisy signal from 65% to 90%. The maximum value for the baseline system with the clean signal was 75%, and the maximum value with the noise signal was 70%. Our system improved the maximum value for both the clean signal and the noisy signal to 100%.

8. FUTURE WORK

Our novel speaker identification system showed significant improvements over the baseline system, showing average results close to 100% for the clean signal. Since the average results for a clean signal are very close to 100%, in the future, we can focus our efforts on improving the results of the system under noisy conditions, where our average result was 90% and our maximum result was 100%. While our alogorithm performed much better than then the baseline system when the noisy signal was used, we will continue to investigate additional modifications that can make the performance of the system under noisy conditions even better.

9. ACKNOWLEDGEMENTS

Our group gratefully acknowledges the support of the UCLA Speech Processing and Auditory Perception Laboratory, including Professor Abeer Alwan and Jinxi Guo, as well as the contributions of the Lawrence Rabiner, Ronald Schafer, and Andrew Moore.

References

[1] Todor Dimitrov Ganchev. *Speaker recognition*. PhD thesis, University of Patras, 2005.

[2] M. Jayasheela. Speaker identification using combined mfcc and phase information.

[3] Hassan Ezzaidi and Jean Rouat. Pitch and mfcc dependent gmm models for speaker identification systems. In *Electrical and Computer Engineering, 2004. Canadian Conference on*, volume 1, pages 43–46. IEEE, 2004.

[4] Urmila Shrawankar and Vilas M. Thakare. Techniques for feature extraction in speech recognition system: a comparative study. *arXiv preprint arXiv:1305.1145*, 2013.

[5] P. P. S. Subhashini and Turimerla Pratap. Text-independent speaker recognition using combined lpc and mfc coefficients.

[6] Michael J. Carey, Eluned S. Parris, Harvey Lloyd-Thomas, and Stephen Bennett. Robust prosodic features for speaker identification. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1800–1803. IEEE, 1996.

[7] Iker Luengo, Eva Navas, Inmaculada Hernáez, Jon Sanchez, Ibon Saratxaga, and Iñaki Sainz. Effectiveness of short-term prosodic features for speaker verification. *Procs. The Fundamentals of Verbal and Nonverbal Communication and the Biometrical Issue. Vietri sul Mare, Italia*, 2006.

[8] Xuejing Sun. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–333. IEEE, 2002.

[9] Ching-Tang . T. Hsieh, Eugene Lai, and You-Chuang . C. Wang. Robust speaker identification system based on wavelet transform and gaussian mixture model. *J. Inf. Sci. Eng.*, 19(2):267–282, 2003.

[10] Amita Dev and B. Parmanand. A novel feature extraction technique for speaker identification. *International Journal of Computer Applications*, 16(6):0975–8887, 2011.