Comparison of Frequency Bands in Closed Set Speaker Identification Performance

[+]Özgür Devrim Orman, *Levent Arslan
[+]TÜBİTAK
Ulusal Elektronik ve Kriptoloji Araştırma Enstitüsü, Gebze, 41470 Kocaeli
oorman@mam.gov.tr
*Boğaziçi Üniversitesi
Elektrik-Elektronik Mühendisliği Bölümü, Bebek, 80815 İstanbul
arslanle@boun.edu.tr

Abstract. Lots of words can be said about the importance of speaker identification for people, but no word might be as meaningful as the imagination of a life without having any speaker identification ability. For example, if we can not identify people from their voices, without having any additional information it is impossible for us to decide on whom we are talking to on telephone. Of course, this ability seems so simple for us, but computer based implementations are still far from human abilities. Furthermore, any speaker identification system on computers can not be designed as an optimum solution. It is known that there is no optimum feature set definition for speaker identification systems. In this work, we study speaker identification performance dependency on the choice of frequency bands.

1 Introduction

Speaker identification process can be subdivided into three phases: i) Transformation of training set speaker records to feature vectors database, ii) Training of the system using these data, and iii) Identification performance test. In the first phase, we can use various methods to generate feature sets, such as LPC cepstrum [1] or mel–cepstrum [2] representations. The process in the second phase depends on the choice of identification method. In this phase we can use Vector Quantisation [3], Gaussian Mixture Models (GMM) [4], Hidden Markov Models [5] or various types of Neural Network architectures such as Radial Basis Function Networks [6,7]. The theoretical details of GMM method are given in Section 2. In the last phase, speaker identification performance of the system is tested using test feature vectors database.

Selection of feature vector parameters has been studied in previous works [1,8,9]. In Sambur's paper [8] important characteristics of various acoustic features are analyzed. These acoustic features are vowels, nasals, strident consonants, fundamental frequency, and timing measurements. Moreover, to determine the overall feature ranking he uses a "knock out" procedure that determines the least important feature parameter at each step using error performance criteria. In Atal's work [1], acoustic parameters in speaker identification are classified in eight different groups. These groups are: intensity, pitch, short–time spectrum, predictor coefficients, formant frequencies and bandwidths, nasal coarticulation, spectral correlations, timing and speaking rate. On the other hand, In O'Shaughnessy's work [9] acoustic features are subdivided into two groups, inherent features and learned ones. F–ratio is accepted as a good measure of the amount of speaker identification information that is carried by any analyzed feature.

Our approach to the feature selection problem differs in many ways from the previous works those we mention. In order to analyze the speaker identification performance dependency on a frequency band, we use training and test sets which are composed of including only the filtered power spectrum values in the analysis frequency range. Besides that, in this work we propose new performance measures which are vector and speaker ranking. The experimental results on speaker identification performance dependency on frequency bands and the methodology are given in Section 3. The results of this work are discussed in Section 4.

2 GMM Based Speaker Identification System

The main idea behind this method is to model the statistical behavior of a speaker's acoustic characteristics by using a mixture of multidimensional Gaussian distributions. Properties of these multidimensional Gaussians, such as mean vectors and covariance matrices, are calculated using Expectation Maximization (EM) algorithm. In this method each speaker is represented by K multidimensional Gaussians. Parameter set of i^{th} speaker is represented as follows.

$$\lambda_i = \{\, p_j, \underline{\mu}_j, \Sigma_j \,\} \quad j = 1,\ldots,K \tag{2.1}$$

$\underline{\mu}_j$: Mean vector of j^{th} Gaussian,

Σ_j : Covariance matrix of j^{th} Gaussian,

p_j : Probability of j^{th} Gaussian.

Conditional probability of observation of the test vector $\underline{x}$ in terms of i^{th} speaker's parameter set is calculated as given below.

$$p(\underline{x}/\lambda_i) = \sum_{j=1}^{K} p_j b_j(\underline{x}) \tag{2.2}$$

$$b_j(\underline{x}) = \frac{1}{(2\pi)^{M/2} \left|\Sigma_j\right|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}-\underline{\mu}_j)\Sigma_j^{-1}(\underline{x}-\underline{\mu}_j)^{\mathrm{T}}\} \tag{2.3}$$

EM algorithm can be formulated as follows.

$$p\left(r/\underline{x}_{i,j,T}\right) = \frac{p_r b_r\left(\underline{x}_{i,j,T}\right)}{\sum_{k=1}^{K} p_k b_k\left(\underline{x}_{i,j,T}\right)} \quad r = 1,\ldots,K \tag{2.4}$$

$$\underline{\mu}_k = \frac{\sum_{j=1}^{D} p(k/\underline{x}_{i,j,T}, \lambda_i)\underline{x}_{i,j,T}}{\sum_{j=1}^{D} p(k/\underline{x}_{i,j,T}, \lambda_i)} \tag{2.5}$$

$$\sigma_k^2 = \frac{\sum_{j=1}^{D} p(k/\underline{x}_{i,j,T}, \lambda_i)x_{i,j,T}^2}{\sum_{j=1}^{D} p(k/\underline{x}_{i,j,T}, \lambda_i)} \tag{2.6}$$

$$p_k = \frac{1}{D}\sum_{j=1}^{D} p(k/\underline{x}_{i,j,T}, \lambda_i) \tag{2.7}$$

In these formulas, $\underline{x}_{i,j,T}$ represents i^{th} speaker's j^{th} training future vector. This optimization procedure is ended, if the calculated likelihood value does not increase more than a predefined threshold between consecutive iterations.

Identification test of any speaker, who is in the set, includes two phases. In the first phase, likelihood value of subject speaker's test set is calculated for each candidate speaker. The second phase includes assignment of speaker who has the highest likelihood ratio, to the subject speaker's identity. Suppose that H represents the assigned speaker and X_S represents the whole set of test vectors of the subject speaker, we can formulate this decision process as;

$$H = \underset{1\le i\le I}{\arg\max}\ \Pr(\lambda_i / X_S) \tag{2.8}$$

Using Bayes rule we can rewrite $\Pr(\lambda_i / X_S)$ as in (2.9).

$$\Pr(\lambda_i / X_S) = \frac{p(X_S/\lambda_i)\Pr(\lambda_i)}{p(X_S)} \tag{2.9}$$

Assuming the probability of each speaker is equal and *$p(X_S)$* value is the same for each speaker, we can simplify (2.9) in (2.10).

$$H = \underset{1\le i\le I}{\arg\max}\ p(X_s / \lambda_i) \tag{2.10}$$

3 Speaker Identification Performance Analysis

Speaker identification system requires both training and test vector sets for speaker identification process. In order to test the speaker identification performance on a discrete frequency band, the train and test sets are generated including only the filtered power spectrum values in analysis frequency range. It is also observed that, these frequency bands must not be shorter than 500 Hz. In the experiments, we use TIMIT speech corpus [10] that has eight different dialect regions of American

English. TIMIT already includes voice active regions in utterances, so in this work we do not need to use a voice activity detection mechanism. The speaker sets we use are restricted only the records of speakers in the fifth dialect region; this approach cancels the effect of dialect region difference in speaker identification performance. Moreover, we work on three speaker sets. First set includes only male speakers, second set includes only female speakers, and third set includes both male and female speakers. The number of speakers in all these sets is equal to twenty-four. Performance analysis in the same gender also eliminates the information carried by gender difference that is valuable for speaker identification. We generate the training set using the unique utterances from all speakers' records, these files have "sa" prefixes, and the files with "si" prefix are used in the test set. Furthermore, phonetic dominance problem in training is cancelled by using these unique utterances.

In the experiments, speech records are segmented in 20 ms frames and the duration between adjacent frames is kept at 10 ms. Each frame is weighted using Hamming window and transformed to frequency domain using DFT, then the power spectrum of a frame is calculated using these coefficients. The power spectrum coefficients are passed through a filter bank that is composed of uniform triangular filters. Train and test files for each frequency band are generated using the filtered power spectrum. The training phase is the same as given in section two. On the other hand, speaker identification performance is measured according to two criteria: vector ranking and speaker ranking.

In *vector ranking*, we compare the statistical likelihood values of each test vector in terms of candidate speakers, and then we assign a rational number between 0 and 1 to the identification performance of the correct speaker. The mean value of all speakers' performance values is calculated and assigned as a final measure of speaker identification performance value for this frequency interval. On the other hand, in *speaker ranking*, we compare the statistical likelihood values of each speaker's test set in terms of candidate speakers, then we do the same numerical assignment as in the previous method that we explain. Also the final measure of speaker identification performance value at this frequency interval according to the speaker ranking is obtained by calculating the average of all speaker's performance values. After we calculate the performance on each frequency band, we can visualize how the speaker identification performance varies along the whole frequency axis. These results are also examined comparing with calculated F–ratio [1] values at each frequency band. Also, F–ratio for this case is the ratio of inter speaker variance to intra speaker variance at that frequency band, and it is interesting to note that there is a correlation between calculated F–ratio values and vector ranking results.

4 Conclusion

Observations in this work give us a new perspective about the importance of frequency bands in speaker identification systems. Although, mel–scale is used in speaker identification systems generally, it is possible to define a new scale using the results of this work. Besides that, we have already developed a new filter bank according the results of this work, it is called as "speaker sensitive frequency scale filter bank" (SSFSF). In the speaker identification test including 462 speakers of

TIMIT corpus, the system with SSFSF gives better identification results as compared with the system including mel–scale filter bank. Furthermore, the following work that we focus on is a subjective test to compare our observations and human auditory system responses.

References

1. Atal, B.S., “Automatic recognition of speakers from their voices”, Proc. IEEE, Vol. 64, pp. 460-474, 1976.
2. Davis, S.B. and Mermelstein, P., “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-28, pp. 357-366, 1980.
3. Rosenberg , A.E., and Soong , F.K., “Evaluation of a vector quantization talker recognition system in text independent and text dependent modes”, Computer Speech and Language, Vol. 22, pp. 143-157, 1987.
4. Reynolds, D.A., Rose, R.C., “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE Trans. Speech and Audio Processing, Vol. 3, pp. 72-83, 1995.
5. Tishby, N.Z., “On the application of mixture AR hidden Markov models to text independent speaker recognition”, IEEE Trans. Signal Processing, Vol. 39, pp. 563-570, 1991.
6. Oglesby, J. and Mason, J., “Radial basis function networks for speaker recognition”, in Proc. ICASSP, May 1991, pp. 393-396.
7. Orman, Ö.D., Arslan L., “A comparative study on closed set speaker identification using RBF network and modular networks”, Accepted for presentation in TAINN’2000.
8. M. R. Sambur, “Selection of Acoustic Features for Speaker Identification”, IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-23, pp. 176-182, 1975.
9. D. O’Shaughnessy, “Speaker Recognition”, IEEE ASSP Magazine, pp. 4-17, October 1986.
10. “Getting started with darpa TIMIT CD-ROM: an acoustic phonetic continuous speech database”, National Institute of Standarts and Technology (NIST), Gaithersburg, MD (prototype as of Dec. 1988).