

Frequency Analysis of Speaker Identification

Özgür Devrim Orman^{1,2}, Levent M. Arslan²

¹TÜBİTAK

National Research Institute of Electronics and Cryptology, Gebze, 41470 Kocaeli

oorman@uekae.tubitak.gov.tr

²Boğaziçi University

Electrical and Electronics Engineering Department, Bebek, 80815 İstanbul

arslanle@boun.edu.tr

Abstract

Our main motivation in this work is to investigate subbands based representation of speaker identities and then search for improvements to typical methods (such as MFCCs). Test results obtained via proposed Vector Ranking criteria, have shown that 0–1000 Hz and 3000–4500 Hz frequency bands are more significant in automatic speaker discrimination when compared to other frequencies. We propose a new filter bank for speaker identification systems. Performances of new cepstras and MFCCs are compared which show that the proposed feature sets' results in significantly better SI performance.

1. Introduction

In general sense, Speaker Identification (SI) problem corresponds to determination of presented talker's identity in a set of speakers. As in the case of all other identification applications, a SI system includes two principal building blocks; feature extraction part which involves a mapping from presented input speech to feature space, and the other one is classifier part which can also be represented as a mapping from feature space to a less dimensional decision space. System performance is of course highly related with the characteristics of both mappings. Two possible ways to increase the SI performance can be listed as; choice of more complicated classification method (such as fusion of multi classifier outputs) and selection of more distinctive acoustic features to represent speaker identity. In this work we are searched for if there exists a better subbands representation of speaker than currently accepted ones, such as MFCC's. As is the case, Speaker Identification Performances (SIP's) on different frequency bands are measured using proposed Vector Ranking (VR) method and scaled in Identification Performance Index (IPI) unit, these results are also compared with calculated F-ratio values of subbands. Furthermore, a filter bank which satisfies obtained IPI results, is designed via a heuristic way.

Which features in speech are appropriate for Speaker Identification and how can they be determined? These questions have been taken into account since the early phonetic studies of this research area [1-3]. Feature extraction approaches employed in the systems participated to NIST 1998 evaluations can be given as current answers for these questions in the literature [4]. Even there are some different choices in acoustic feature extraction methods (such as nonlinear discriminant analysis, pitch prosodics, modulation

spectral filtering, and speaker mapping), majority of systems in the evaluations include subband based or LPC derived cepstrum calculations. Moreover, discriminant frequency bands in SI were also studied in some previous works [5-6], and considering performance definition differences between our approach and these studies, it is quite interesting to note that some results are highly correlated with what we obtained.

Details of this work are presented in the following order. Methods used in SIP analysis are presented in Section 2. Developed SIP tests are given in Section 3. Proposed Filter Bank (PFB) and comparison of SI results of both PFB Cepstrum Coefficients (PFB-CCs) based and MFCCs based systems are explained in Section 4. Conclusions of this study are given in Section 5.

2. Methods for SIP Analysis

The first type of method explained here is based on calculation of speaker identification performances on different frequency bands using vector ranking (VR), and results for this case are given in a unit called as Identification Performance Index (IPI). Very early studies on both VR and IPI were presented in [7]. The other type of method includes calculation of F-ratio values on subbands. F-ratio has also been used in previous works to evaluate significance of acoustic features [1-3].

Analytic formulation of VR criteria is explained as: Let $V_{m,n}\{ \cdot \}$ operator determines the rank order of m^{th} speaker's likelihood value in likelihood values of all speakers for n^{th} token of m^{th} speaker (for example, if there are 24 speakers, and the likelihood of the correct speaker for presented acoustic vector was the 3rd highest value, then the VR result was 3/24).

$$V_{m,n}\{l(x_{m,n}/A_1), l(x_{m,n}/A_2), \dots, l(x_{m,n}/A_M)\} = \frac{R_{m,n}}{M} \quad (1)$$

In Equation (1) A_m represents m^{th} speaker's GMM classifier parameters (mean vectors, covariance matrices and component densities), $x_{m,n}$ represents m^{th} speaker's n^{th} token, $l(x_{m,n}/A_j)$ corresponds to likelihood value of m^{th} speaker's n^{th} token belongs to j^{th} speaker, $R_{m,n}$ means likelihood rank order of m^{th} speaker for his n^{th} token, and M is the number of speakers.

$$IPI = \frac{1}{M \cdot N} \sum_{m=1}^M \sum_{n=1}^N V_{m,n} \{l(x_{m,n}/A_1), \dots, l(x_{m,n}/A_M)\} \quad (2)$$

Right hand side of Equation (2) includes two averaging operations: the first one is averaging of VR values of each speaker over all his tokens (each speaker has N tokens) and the other one is averaging of values obtained from the first operation over all speakers.

F-ratio value of each frequency band in the test corresponds to the ratio of inter-speaker to intra-speaker variance at that band, and Atal's work [3] is a good reference for F-ratio method. Values obtained from the F-ratio test might be expected to represent effectiveness of each frequency band in speaker discrimination.

It is interesting to note that, a strong correlation between the results of both methods is experimentally observed.

3. SIP Analysis on Subbands

Three speaker sets (males, females and the one includes equal representation of both genders) including 24 speakers are employed in the test, and in order to cancel out the effect of dialect region differences they are restricted to have only the speakers from the fifth dialect region of TIMIT corpus [8]. Expected benefit of conducted performance analysis in the same gender is the cancellation of gender difference information that may affect SIP. The reason behind all these restrictions is that, we are searching for a subbands based representation which is able to discriminate between speakers with "similar" acoustic characteristics. Moreover, the training set is generated using the files with "sa" prefixes which are common across all speakers, and the files with "si" prefix are used in the test set, by this way possible dominance of any phoneme to the others in training is omitted.

We can subdivide the feature extraction process into four phases. In the first phase, speech records are segmented to 20 ms frames with 50% overlap. In the second phase, frames are weighted using a Hamming window and transformed into the frequency domain via DFT. In the third phase, power spectrum of each frame is calculated. Then, power spectrum coefficients are passed through a filter bank that is composed of 64 uniform triangular filters with 50% overlap. After passing the power spectrum through the filter bank, each analysis frequency band corresponds to four coefficients. Training and test files are composed of these four-dimensional vectors. It is also experimentally observed that, the length of these frequency bands must not be shorter than 500 Hz. Using generated feature sets, speaker identification performance is measured according VR.

In the test procedure for each frequency band, GMM based SI system includes 32 different Gaussian components the calculation of feature vectors should represent the importance of this range. Except of females, speaker identification performance in the other two figures slightly decreases around frequencies beyond 4500 Hz. The minor

for each speaker is trained, then speakers' likelihood values for each test vector of the specified frequency band are calculated, and finally IPI value of each frequency is obtained via VR method.

Obtained IPIs for the speaker set which includes equal number of males and females are shown in Figure 1. IPI values for males and females are given together in Figure 2, in which dark bars correspond males' scores and the lighter bars represent females' data.

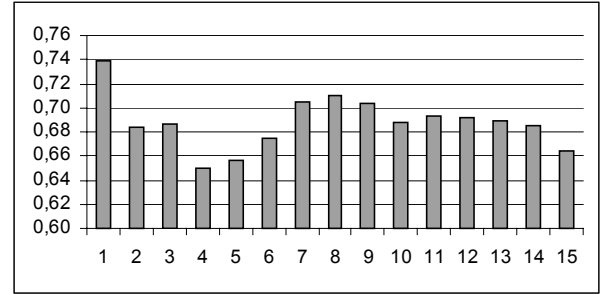


Figure 1: IPI values for both genders.

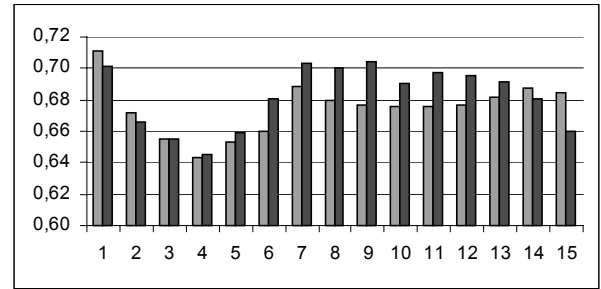


Figure 2: VR results for males and females.

If we look at Figure 1, we see that, the first frequency band gives the highest speaker identification performance on this speaker set. Moreover, if we look at Figure 2, observed performance characteristics on this band are approximately the same as the ones in Figure 1. Since the frequency range of this band includes the first formant of human speech, observed results may be explained as the importance of the first formant for identification systems. In these two figures, speaker identification performance generally decreases between 1 kHz and 3 kHz, so the importance of this range for identification systems can be defined as low. Although the results indicated in these figures are not quite the same, we can say that the identification performance is better between 3000 Hz and 4500 Hz frequencies. The filter bank that is used in

increase of identification performance at high frequencies among female speakers might be explained to some extent with the knowledge that female speech has higher formant frequencies when compared to male speech. Using this

information and simulation results shown in Figure 2, we can say that high frequency bands in female speech carry valuable speaker discriminative information.

In Figure 3, calculated F-ratio values of the speaker set includes both genders (the same as the one in Figure 1) are shown. The highest F-ratio value occurs in the first frequency band. We also note a rise in F-ratio value around mid-frequencies. Figure 4, which shows the F-ratio values obtained from females and males are quite similar to Figure 3, in this figure dark bars are males' results and the lighter bars represent females' scores. The rise in the mid-frequency region for male speakers is more pronounced when compared to other sets. According to these observations, it is important to note that there is a strong correlation between the results of F-ratio test and VR based IPI values. It is the significance of low frequencies and mid-frequencies.

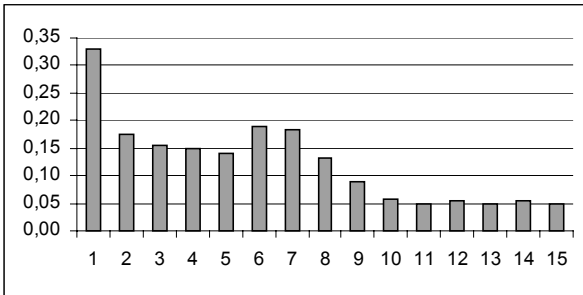


Figure 3: F-ratio values for speaker set of both genders.

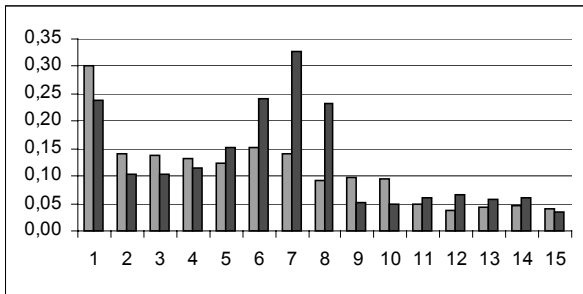


Figure 4: F-ratio values for males and females.

4. Proposed Filter Bank

Based on test results on various frequency bands, we have observed that speaker ID systems show different responses in different frequency bands when compared to speech recognition systems. Therefore one may argue that the acoustic feature set which is optimal for speech recognition task (such as MFCCs) might not be optimal for speaker

5. Conclusions

Observations in this study give new facts about the importance of different subbands for SI systems. Presented contributions are: detailed investigation of SI performances on various frequency bands, definition of a new speaker

identification. One of the contributions of this work is the PFB that is developed heuristically according to the IPI results. PFB is composed of triangular filters like mel-scale filter bank and both filterbanks with 24 subbands are shown in Figure 5.

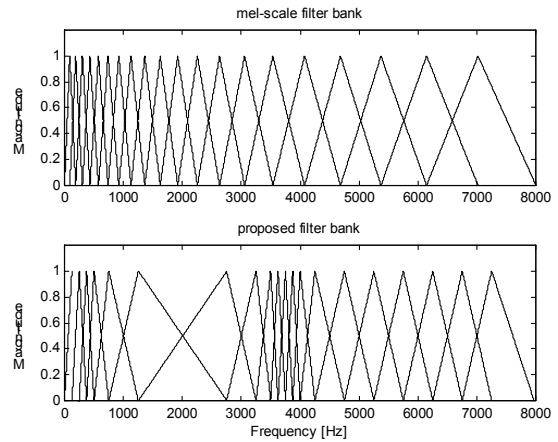


Figure 5: Proposed filter bank and mel-scale filter bank.

In order to make the performance comparison, two different groups of training and test sets are built. One of them includes MFCCs and another includes PFB-CCs. Except of the use of different filter banks each consist of twenty triangular filters, extraction processes of both features are similar. Speaker set includes 462 speakers is composed of both genders from eight different dialect regions of TIMIT corpus. Speaker models are trained using the files with "sa" prefix, and the files with "si" prefix are used in the test.

GMM based speaker identification system is used in which each speaker is represented by 32 Gaussian components with diagonal covariance matrices. Speaker models are trained using EM algorithm [9]. Final speaker decision includes the comparison of each candidate speaker's likelihood values for presented test set and the speaker with maximum likelihood value is identified.

Using mel-scale filter bank based speaker identification system, 385 of 462 speakers are identified correctly. On the other hand, PFB-CCs based system identifies 413 of 462 speakers correctly, which is greater than mel-scale based method. These results are lower than the ones in the literature [9], because the amount of training data for this case is considerably limited as compared with the other works. Besides, our main focus in this paper is to propose better feature extraction methods rather than proposing better modeling methods.

identification performance measure (VR), and finally proposal of a new filter bank for SI.

Obtained test results indicate that, if an SI implementation includes perceptually motivated feature extraction processes, the importance of mid and low frequencies should be considered in the design phase of the filter bank. In this work we propose a new filter bank for speaker identification that is

designed by considering the effectiveness of different frequency bands in IPI. A speaker identification test including 462 speakers of TIMIT corpus was done and the system with PFB gives better identification result as compared with the system including mel-scale filter bank.

A new frequency scale definition for speaker identification can be completed in future works, which requires more detailed investigations via objective tests on various speaker sets. The performance of the proposed scale must be tested for different modeling methods and various languages as well.

One another future study might be on fusion of SI results on single subbands or subband sets which are composed by considering their correlation in SI.

6. References

- [1] Sambur, M. R., "Selection of Acoustic Features for Speaker Identification", *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-23, 176-182, 1975.
- [2] Wolf, J. J., "Efficient Acoustic Parameters for Speaker Recognition", *J. Acoust. Soc. Amer.*, 51, 2044-2056, 1972.
- [3] Atal, B. S., "Automatic recognition of speakers from their voices", *Proc. IEEE*, 64, 460-474, 1976.
- [4] Doddington, G. R. et al., "The NIST speaker recognition evaluation-Overview, methodology, systems, results, perspective", *Speech Com.*, 31, 225-254, 2000.
- [5] Besacier, L., Bonastre, J. F., Fredouille, C., "Localization and selection of speaker-specific information with statistical modeling", *Speech Com.*, 31, 89-106, 2000.
- [6] Besacier, L., Bonastre, J. F., "Subband approach for automatic speaker recognition: optimal division of the frequency domain", In: Bigün et al. (Eds.), *Proc. of the Audio and Video based Biometric Person Authentication, Springer LNCS, New York*, 195-202, 1997.
- [7] Orman, Ö.D., Arslan, L.M., "Comparison of Frequency Bands in Closed Set Speaker Identification Performance", In: Sojka P. et al. (Eds.), *Proc. of TSD'2000, Springer LNAI, Brno, Czech Republic*, 314-318, 2000.
- [8] Fisher, W. et al., "An acoustic-phonetic database", *J. Acoust. Soc. Amer.*, suppl A., 81(S92), 1986.
- [9] Reynolds, D. A., Rose, R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech and Audio Proc.*, 3, 72-83, 1995.