Semi-automatic Speaker Verification System Based on Analysis of Formant, Durational and Pitch Characteristics

Elena Bulgakova[1,2(✉)] and Aleksey Sholohov[1]

[1] ITMO University, St. Petersburg, Russia
{bulgakova,sholohov}@speechpro.com
[2] Speech Technology Center, St. Petersburg, Russia

Abstract. Modern speaker verification systems take advantage of a number of complementary base classifiers by fusing them to get reliable verification decisions. The paper presents a semi-automatic speaker verification system based on fusion of formant frequencies, phone durations and pitch characteristics. Experimental results demonstrate that combination of these characteristics improves speaker verification performance. For improved and cost-effective performance of the pitch subsystem further we selected the most informative pitch characteristics.

Keywords: Formant frequencies · Phone durations · Pitch characteristics · Speaker verification · Feature selection

1 Introduction

Speech signals carry different information including individual voice characteristics which allows to recognize people by their voice, and therefore to solve a speaker recognition task. This task involves speaker verification in case it is necessary to make a binary (yes or no) decision regarding speaker identity, and speaker identification in case it is necessary to determine which speaker voice is presented on a test recording. In this study we focus on a speaker verification problem. Nowadays human-assisted methods are widely used in forensic speaker recognition [1]. However, the application of these methods is limited by the need of engagement of highly qualified experts. Moreover, human-assisted methods are time consuming that generally complicates their use under time constraints. Furthermore, the final decision is largely subjective since it depends on the personal opinion of the expert [2]. In this paper we continue our research started in [3] and propose a semi-automatic speaker verification system which makes it possible to get over above-mentioned shortcomings. This system includes comparing different voice characteristics: formant frequencies, phone durations and pitch characteristics as well. The final decision concerning identity or difference of speaker voices is made automatically as a result of fusion of the used subsystems. The results of our experiments show that additional use of the pitch

A. Ronzhin et al. (Eds.): SPECOM 2016, LNAI 9811, pp. 613–619, 2016.
DOI: 10.1007/978-3-319-43958-7_74

subsystem proposed in [4] leads to better performance compared with the results of our previous research [3]. For the purpose of increasing verification accuracy and time reduction of comparing speech samples, we found the most distinctive pitch characteristics.

The rest of the paper is organized as follows. Section 2 includes the system description. The experimental results and database descriptions are presented in Sect. 3. Conclusions are considered in Sect. 4.

2 System Description

The proposed speaker verification system consists of three subsystems based on pitch characteristics, formant frequencies and phone durations described in Sects. 2.1, 2.2 and 2.3. Figure 1 shows main modules of the system. The first module in each subsystem extracts speech features from the input speech signal. The second module aggregates these features to represent an entire utterance as a vector of fixed dimension. Given a trial each subsystem outputs a matching score measuring similarity between two utterances. At the fusion stage matching scores are combined into a single score to increase accuracy of the system. The decision module compares the final matched score to a pre-defined threshold. If similarity is above the threshold, the trial is classified as target, otherwise non-target.

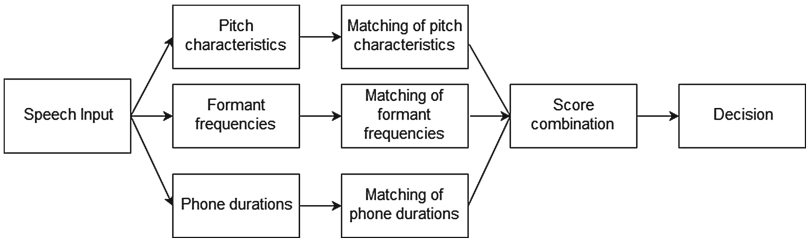


Fig. 1. Block diagram of the speaker verification system

2.1 Pitch Subsystem and Pitch Characteristics

The pitch subsystem compares the characteristics of intonation structures presented in speech samples [4]. The following characteristics were used and described in [4]: initial, final, minimal, maximum and average frequencies of intonation fragments, F0 range, pitch change speed, irregularity coefficient, skewness and kurtosis of distribution of pitch frequencies, duration of intonation fragments, coordinate of minimal, average and maximum frequency values (in percentage of whole duration of chosen fragments). Data analysis includes calculating and correcting pitch curves, segmentation of speech material into

comparable intonation structures (fragments) of the utterances (prosodic phrase, head, pre-head, nuclear tone, nucleus + tail) and automatic comparison of pitch characteristics obtained as a result of segmentation. Because of the high labour intensity of this method we conducted segmentation of speech material based on prosodic phrases of 10–15 s duration in an automatic mode without preliminary pitch correction.

2.2 Formant Subsystem and Formant Frequencies

It is well-known that positions of the main spectral peaks in the spectrum of the speech signal depend on the anatomical structure of the vocal tract and the sizes of the resonant cavities. For this reason such spectral characteristics may be applicable to speaker recognition. Since formant frequencies are usually not independent, we use a GMM-UBM framework [5] which is a common tool in speaker verification to approximate complex statistical relationships in multivariate data. It is based on the notion of the universal background model (UBM) which models statistical distribution of features for a large population of speakers [6].

It should be noticed that hand-correcting formant tracks was not carried out. For our experiments we detected the first four formant tracks of six Russian vowels (/i/, /e/, /a/, /u/, /o/, /y/).

2.3 Phone Subsystem and Phone Durations

The phone duration subsystem was presented in [3]. This subsystem includes automatic phonetic segmentation on the basis of recordings and text contents of these files, calculation of average durations for each phone in the phonetic segmentation and calculation of the matching score of speaker voices. Unlike the formant subsystem based on the GMM-UBM framework which enjoys large speech datasets, training the phone subsystem requires transcriptions (typically limited) in addition to speech recordings. Thus smaller amounts of data may lead to over-fitting because of a large number of model parameters. Due to the lack of text contents of speech recordings, we define a simple matching score which has much smaller parameters to tune and hence more robust to over-fitting:

$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = -\sum_{t=1}^{T} w_i (x_1^t - x_2^t)^2, \tag{1}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2$ is a pair of feature vectors representing a trial, T is the feature space dimension and w_i are non-negative weights. This formula can be seen as negative Mahalanobis distance. Intuitively greater weights should correspond to more important features (*i.e.* features with higher discriminative ability). We give details how to estimate these weights in the next Section. To the aim of time reduction we did not correct phone boundaries.

3 Experiments

3.1 Experiment – Speaker Verification

Here we describe the experiment on speaker verification. For the experiment presented in this Section we used the database described below. For training we formed the database including Russian quasi-spontaneous speech of 124 male speakers and 70 female speakers recorded over the telephone channel. Each speaker participates in five recording sessions of 3–5 min duration and there is one week gap between sessions. During the recording session every speaker answers questionnaire questions. For training we used the database of 1–3 min natural spontaneous telephone dialogues in Russian. The evaluation set consists of 1037 target and 9397 non-target trials for males and 507 target and 2233 non-target trials for females. To increase reliability of speaker verification the final decision can be made based on decisions of independent subsystems. Such procedure is called a decision fusion at the score-level [7]. For a set of matching scores s_i fusion was done using a convex combination of scores:

$$s = p_1 s_1 + p_2 s_2 + p_3 s_3,$$

where s_i is a matching score of the i-th sub-system, p_i are weight parameters such that $\sum_i p_i = 1$. The values p_i were tuned by hand on a subset of the training set.

The important aspect of fusion is statistical independence of matching scores of combined subsystems. Otherwise the final decision hardly results in a sharp gain in speaker verification performance.

We report speaker verification performance in the form of equal error rate (EER, %) [8]. Table 1 presents the performance evaluation of the considered subsystems.

Table 1. Speaker verification results for two different genders (EER, %)

Subsystem	Male	Female
Pitch characteristics	23.28	27.33
Phone durations	27.57	36.98
Formant frequencies	2.93	4.63
Formant frequencies + Phone durations	2.02	4.49
Formant frequencies + Phone durations + Pitch characteristics	1.41	3.83

As appears from Table 1, the formant subsystem is the most accurate. Pitch characteristics demonstrate the noticeable degradation. Phone durations concede in performance to other characteristics. In our previous research fusion of subsystems based on phone durations and formant frequencies was performed [3]. In this experiment we conducted fusion of all above-described subsystems.

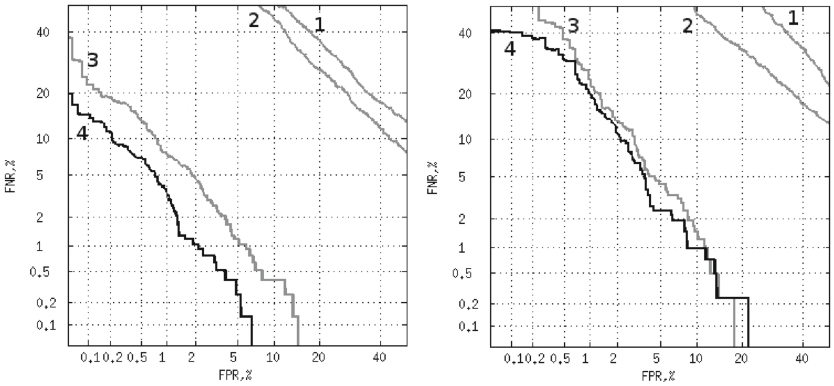


Fig. 2. DET (detection error trade-off)-curves for male (a) and female speakers (b). DET-curve (1) demonstrates the performance of the phone durations subsystem, (2) – pitch subsystem, (3) – formant subsystem, DET-curve (4) shows the whole system performance. FNR (False Negative Rate), FPR (False Positive Rate).

The results presented in Table 1 and Fig. 2 demonstrate that fusion of subsystems based on poorly correlated features (pitch characteristics, formant frequencies and phone durations) leads to a decrease of EER and improves speaker verification performance.

3.2 Experiment – Informative Pitch Characteristics

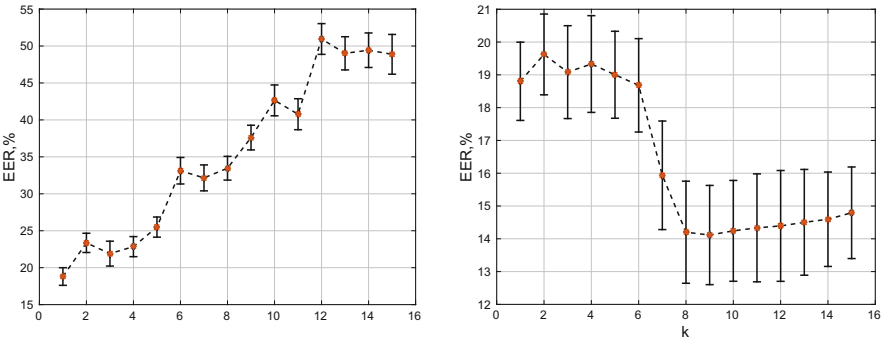
Feature selection is the crucial step in design semi-automatic speaker recognition systems. It can considerably reduce time of comparing speech samples and even improve speaker verification performance.

We ranked features according to weights calculated as follows:

$$w_i = \frac{\sigma_b^2}{\sigma_w^2}, \tag{2}$$

where σ_b^2 is between-speaker variance and σ_w^2 is within-speaker variance for the i-th feature. Higher values correspond to features with a higher class separability. To assess selected subsets of features we evaluated speaker verification performance as the function of a number of selected features. We used a dataset consisting of 5102 speech cuts from 195 speakers. Each speaker takes part in 2–5 recording sessions of 3–5 min duration. The database includes male and female spontaneous speech of speakers recorded over a microphone channel in Russian, Tajik, Azerbaijani and Talysh. It should be noticed that prosodic segmentation into intonation fragments was done fully manually. To evaluate verification performance we averaged EERs over 100 random splits of the dataset into equally-sized training and testing parts. First, we estimated system accuracy

in terms of EER using subsets of the most informative features. Starting from the top ranked feature we gradually added other features according to the order defined by weights (2). We used the same weights to compute the matching score defined by (1). Then we estimated EERs for each feature separately. Figure 3 demonstrates the results.



(a) (b)

Fig. 3. Speaker verification performance (a) with a subset of the top-k most informative pitch characteristics and (b) for each characteristic separately (sorted in the same order).

Figure 3(a) shows individual speaker verification performance for all pitch characteristics ordered according to (2): (1) average, (2) final, (3) minimal, (4) initial, (5) maximum pitch frequencies, (6) F0 range measured in Hertz, (7) irregularity coefficient, (8) pitch change speed, (9) F0 range measured in semitones, (10) kurtosis and (11) skewness of distribution of pitch frequencies, coordinate of (12) minimal, (13) maximum and (14) average frequency values (in percentage of whole duration of chosen fragments), (15) duration of intonation fragments. Thus the first five most informative features are (1–5), while (12–15) are the least distinctive features. However, as can be observed, there is a strong correlation between some features. For this reason the joint use of such characteristics as (1–5) does not improve speaker verification performance that Fig. 3(b) shows. While adding (7) and (8) leads to a noticeable decrease of EER. Interestingly, including the rest of less informative pitch characteristics even slightly decreases accuracy of the pitch subsystem. The results of the additional experiments demonstrate that the joint use of (1), (7) and (8) leads to the best speaker verification performance having the lowest EER of 13 %. Therefore, EER obtained on the reduced feature set is lower than that on the full feature set (14,79 %). It was also experimentally established that the threshold for absolute difference of average pitch frequencies corresponding to equal misses and false alarms equals to 12 Hz. This finding can be useful for rapid comparison of speech samples carried out by experts.

4 Conclusion

In this paper we proposed a semi-automatic speaker verification system based on fusion of formant frequencies, phone durations and pitch characteristics. Experimental results show that including of pitch characteristics improves speaker verification performance compared with an earlier developed system [3]. We found out that use of the reduced set of pitch characteristics (average F0, irregularity coefficient and pitch change speed) leads to increased speaker verification accuracy.

Acknowledgments. This work was financially supported by the Government of the Russian Federation, Grant 074-U01.

References

1. Rose, P.: Forensic Speaker Identification. Taylor and Francis, London (2002)
2. Tanner, D.C., Tanner, M.E.: Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection. Lawyers and Judges Publishing, Tucson (2004)
3. Bulgakova, E., Sholohov, A., Tomashenko, N., Matveev, Y.: Speaker verification using spectral and durational segmental characteristics. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS, vol. 9319, pp. 397–404. Springer, Heidelberg (2015)
4. Smirnova, N., et al.: Using parameters of identical pitch contour elements for speaker discrimination. In: Proceedings of the 12th International Conference on Speech and Computer, pp. 361–366 (2007)
5. Becker, T., Jessen, M., Grigoras, C.: Forensic speaker verification using formant features and Gaussian mixture models. In: Proceedings of Interspeech, pp. 1505–1508 (2008)
6. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digit. Signal Proc. **10**, 19–41 (2000)
7. Jain, A.K., Flynn, P., Ross, A.A. (eds.): Handbook of Biometrics. Springer-Verlag New York, Inc., New York (2008)
8. The NIST year 2010 Speaker Recognition Evaluation plan. http://www.itl.nist.gov/iad/mig/tests/sre/2010/NISTSRE10evalplan.r6.pdf