Speaker Verification Using Spectral and Durational Segmental Characteristics

Elena Bulgakova[1], Aleksei Sholohov[1], Natalia Tomashenko[1,2](✉), and Yuri Matveev[1,2]

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

Abstract. In the present paper we report on some of the results obtained by fusion of human assisted speaker verification methods based on formant features and statistics of phone durations. Our experiments on the database of spontaneous speech demonstrate that using segmental durational characteristics leads to better performance, which shows the applicability of these features for the speaker verification task.

Keywords: Spectral formant features · Segmental durations · Speaker verification

1 Introduction

Information contained in speech signal makes it possible to solve one of the most important problems of modern speech technology - the problem of speaker verification. This task involves comparing test and model recordings to confirm the identity of speakers' voices in the presented recordings. At the present time, automatic and human assisted methods are widely used for solving the problem of speaker verification. While automatic methods usually give superior performance, human assisted methods make it possible to clarify and correct the work of automatic methods. They are also applied in cases where the work of automatic methods is restricted, for example, under high noise conditions.

Earlier studies into forensic speaker recognition which include the speaker verification task are often concerned with the statistical analysis of the distribution of such acoustic and prosodic features as fundamental frequencies [1–3], formant frequencies [4–7] and temporal suprasegmental characteristics [8,9]. Relatively little attention has been given to speaker specific segmental durations. However, such information is valuable in distinguishing speakers [10].

These features are especially useful for verification of speakers with a similar vocal tract, when some other features (*e.g.* spectral characteristics) are not sufficiently trustworthy. In this paper we study the applicability of durational characteristics for the speaker verification task as well as the possibility to use them with the other features. To this aim we implement a fusion of the phone durations method and the method based on formant features [7] and compare the obtained results with the performance of the human assisted pitch method [3].

© Springer International Publishing Switzerland 2015
A. Ronzhin et al. (Eds.): SPECOM 2015, LNAI 9319, pp. 397–404, 2015.
DOI: 10.1007/978-3-319-23132-7_49

2 Speaker Verification Methods

2.1 Formant Method

The main spectral peaks (formants) are influenced by the anatomical structure of the vocal tract and the sizes of the resonant cavities. For this reason such features may be useful for speaker discrimination. We extract the values of the first four formants for 6 Russian vowels (/i/, /e/, /a/, /u/, /o/, /y/).

Since formant values are usually not independent from each other, there is a need for modeling of complex statistical relationships of formants values in speech for each speaker. Currently, one of the most common approaches to modeling complex multivariate distributions for speaker recognition is GMM-UBM framework [7]. The key idea of this approach is to construct so-called *universal background model* (UBM), which approximates the feature distribution of a large number of speakers to represent the whole population. Both UBM and speaker models are implemented by means of a Gaussian mixture model (GMM), which is a weighted sum of K multivariate Gaussian distributions:

$$p(\boldsymbol{x}|\theta) = \sum_{k=1}^{K} \pi_k g(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

where $\boldsymbol{\mu}_k$ is the mean vector and $\boldsymbol{\Sigma}_k$ is the covariance matrix of k-th component, π are mixture weights summing to 1 and the tuple $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ represents all the GMM parameters.

Given a set of feature vectors $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ assumed to be statistically independent and a speaker model θ, the *likelihood* function measures how well the model θ fits the data $\boldsymbol{X}$:

$$P(\boldsymbol{X}|\theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta), \tag{2}$$

where $p(\cdot|\theta)$ is Gaussian mixture density (1) representing the speaker model θ. Given a speaker model $\theta_{speaker}$, UBM – θ_{UBM} and the set of fetures $\boldsymbol{X}_{test}$ extracted from a test recording, the decision of the identity or difference of the two speakers can be made on the basis of the following rule:

$$\frac{P(\boldsymbol{X}|\theta_{speaker})}{P(\boldsymbol{X}|\theta_{UBM})} < \Lambda, \tag{3}$$

where Λ is the decision threshold value set in advance. Since numenator represents the hypothesis that the test feature vectors originate from the model $\theta_{speaker}$, higher likelihood ratio (3) favors this hypothesis. Otherwise, it is more likely that $\boldsymbol{X}_{test}$ comes from different speaker.

In our experiments we used UBM with 16 mixture components.

2.2 Phone Durations Method

The main stages of the algorithm based on statistics of phone durations include:

1. *Automatic phonetic segmentation* on the basis of recordings and text contents of these files. In the course of the segmentation, temporary boundaries of each phone are defined. After carrying out automatic segmentation the expert can correct the boundaries of the allocated phones if necessary.
2. *Calculation of average durations* for each phone in the phonetic segmentation.
3. *Calculation of a matching score* of speakers' voices and decision-making.

We produce forced alignment of speech audio files with its transcription at the phone level. The number of phone classes is 53, they correspond to 52 (17 vowels and 35 consonants) phones of the Russian language and a silence model. The six vowel symbols (i, e, a, u, o, y) acquire a numerical index specifying vowel position in relation to the stressed syllable: "0" denotes a vowel in a stressed syllable, "1" denotes a prestressed vowel (for vowel /a/ – only the 1st pre-stressed syllable or the initial word position), "2" stands for a second pre-stressed position of vowel /a/, while "4" indicates any post-stressed position of all vowels. The 3rd degree of vowel reduction indicated by "3" in some notation systems is excluded from our vowel classification. For producing phonetic segmentation we trained a Hidden Markov Model (HMM) acoustic model on 150 h of audio data from a Russian speech dataset. The training set consists of reading, spontaneous conversational speech and records of TV broadcasts. The acoustic model is a standard tandem GMM-HMM with tied-state context-dependent triphones, where each model, except the silence, has left-to-right, 3-state topology [12]. The silence model has one state. The total number of tied states is 13700 with, on average, 14 Gaussians per state. Acoustic features are LC-RC [11]. In practice, available transcriptions for some speakers may be poor, for example, when they do not correspond exactly to the audio content. A traditional approach to segmentation, such as forced-alignment with Viterbi algorithm, fails to work under these conditions. Hence, we implement a two-stage segmentation algorithm, similar to that, proposed in [13,14]. Figure 1 shows an example of phonetic segmentation.

Thus, unlike the formant-based method, training a statistical model of the speaker's voice requires transcription as well as speech recording. If a sufficiently large number of training files are available we can in principle apply the generic GMM-UBM approach. Otherwise, UBM would significantly differ from the actual statistical distribution of features in the population, leading to very poor performance as a result of overfitting because of a large number of model parameters. In our case, due to lack of transcribed utterances, we define a simple matching score as follows:

$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = -\sum_{t=1}^{T} w_i (x_1^t - x_2^t)^2, \tag{4}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2$ are the vectors of mean durations representing a trial, T is the total number of phones and w_i are the nonnegative weights. These weights should con-

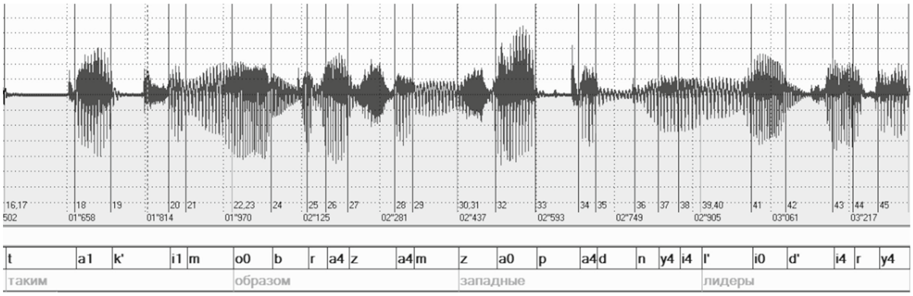


Fig. 1. Screenshot of phonetic segmentation of the utterance "takim obrazom zapadnye lidery"

form to the importance of a phone for speaker discrimination. Following the intuition that phones with greater discriminating ability should have lower *within-speaker variability* and higher *between-speaker variability* at the same time, we define the weights in the following way[1]:

$$w = \frac{\sigma^b}{\sigma^b + \sigma^w}. \tag{5}$$

We define the notion of between- and within-speaker variability as between- $\sigma^b = \frac{1}{S}\sum_s (m^s - m)^2$ and within-speaker $\sigma^w = \frac{1}{N}\sum_s \sum_r (x^{sr} - m^s)^2$ variances respectively, where s runs over all the S speakers in training set, r stands for the session index for each speaker. Here $m^s = \sum_r x^r$ is within-speaker mean, $m = \sum_s m^s$ is the mean of means and N is the training set size.

2.3 Pitch Method

The pitch method allows the expert to analyze and compare the main characteristics of intonation structures presented as sets of pitch parameter values for comparable units of melodic contour. The possibility of comparison of melodic structures is provided with their relative stability within-speaker variability in comparable contexts. Pitch analysis consists of finding of the same intonation structures in the studied recordings and comparing their characteristics. The data analysis includes obtaining and correcting the pitch files for neutral declarative utterances and building the tables containing the values of pitch parameters for the structural intonation units of the utterances (prosodic phrase, head, prehead, nuclear tone, nucleus + tail). Because of the high labour intensity of this method we performed segmentation of speech material based on prosodic phrases of 10–15 s duration. Such procedure is possible as this intonation structure is considered to be the most informative. After that we performed statistical analysis of basic intonation structures used by speakers. The following characteristics were taken as pitch parameters: minimum, maximum and average frequency

[1] Superscript t is omitted for the sake of presentation clarity.

values, F0 interval measured in Hz and semitones, pitch change speed, second irregularity coefficient [3].

3 Experiments

3.1 Database

For training we recorded a database consisting of 194 Russian native speakers. The speech data include quasi-spontaneous Russian speech of 124 male speakers and 70 female speakers recorded through the telephone channel. During the recording each informant answers the questions of the questionnaire. Every speaker takes part in five recording sessions of 3–5 min duration and there is a gap of at least one week between two sessions. For testing we recorded a database of 1–3 min natural spontaneous telephone conversations between two Russian native speakers. This evaluation set consists of 1037 target and 9397 non-target trials for males and 507 target and 2233 non-target trials for females.

3.2 Experiment – Speaker Verification

Here we describe experiments in speaker verification using the database, as described above. We report speaker verification performance in the form of *equal error rate* (EER,%) [15]. In the first experiment, we compared performance of three human assisted methods. Two of them (the formant and phone methods) were compared in a completely automatic mode, *i.e.* without hand-correcting formant tracks and phone boundaries. Trials using the pitch method were conducted in the semi-automatic way as discussed in Sect. 2.3. Because of the high labour intensity of the pitch method, we did not have the possibility to use the full test set. Therefore, we selected a subset consisting of 50 targets and 50 impostors.

Table 1 presents the results of comparison. As can be seen from Table 1, the formant method is the most accurate of all.

Table 1. Results for speaker verification on toy test (EER, %)

Method	EER, %	
	male	female
Pitch	12.5	13.6
Phones	31.3	33.8
Formants	2.0	2.0

To study the possibility of joint use of the compared methods we carried out their fusion. As the pitch method demands considerable time for data preparation, it was excluded from fusion.

For two matching scores s_1 and s_2 fusion was performed at the score-level using simple convex combination with a weight α to get the final score:

$$s = \alpha s_1 + (\alpha - 1)s_2$$

The optimal value of α was found using a subset of training set. It was close to 0.9.

Table 2 shows that fusion of methods based on formant features and statistics of phone durations decreases EER and results in sharp gains in performance. It means that using segmental durational characteristics improves the speaker verification performance.

Table 2. Results for speaker verification (EER, %)

Method	EER,%	
	male	female
Formants	3.2	4.8
Formants+Phones	2.4	4.5

3.3 Experiment – Informative Phones

Here we set the task of finding the most informative phones in terms of their ability of speaker discrimination. As discussed in the previous section, the formula (5) may be an indicator of the discriminative ability. According to this definition of informativity we can list the phones having the largest values of (5): /t/, /n/, /r/, /v/, /p/, /a0/, /l/, /k/, /o0/, /a1/ for females and /l'/, /l/, /ch/, /n'/, /r/, /t'/, /n/, /r'/, /a0/, /a1/ for males. Figure 2 shows weights for the case of female gender.

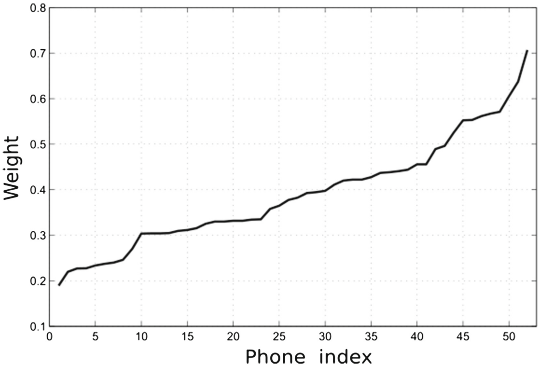
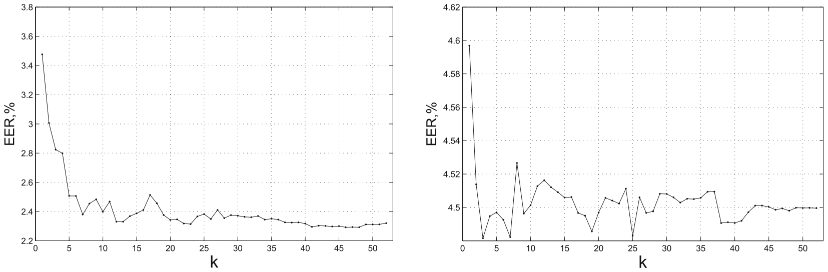


Fig. 2. Weights calculated on female data and sorted in ascending order.

Experimental results allow us to draw the conclusion that the majority of phones found with the greatest weight represent vowels and sonants. Interestingly, such phones as vowel /a/ in the pre-stressed syllable, sonants /n/, /r/, /l/ have the best discriminative ability both for male and female speakers. However, it is possible to note some gender distinctions. For example, vowels, /a/, /o/ in the stressed position as well as voiceless stops /p/, /t/, /k/ and voiced labiodental fricative /v/ are especially important for discrimination of female speakers while affricate /ch/, sonants /n'/, /r'/, /l'/ and voiceless dental stop /t'/ belong to informative phones found for male speakers.

To visualize this method of phone importance ranking, we conducted the following experiment. Starting from the one most informative phone we gradually added one by one all remaining phones ordered according to their relevance for speaker discrimination. At each step k we measured speaker verification performance for the top-k most informative phones. In other words, only a subset of phones was used to compute the sum (4).



(a)

(b)

Fig. 3. Speaker verification performance with a subset of the top-k most informative phones for males (a) and females (b).

We can see that the first few phones make it possible to reach the performance comparable to the best performance using larger subsets of phones. Interestingly, including the rest of less informative phones leads to slight degradation (Fig. 3).

4 Conclusion

In this paper we approved the applicability of segmental durational characteristics for the speaker verification task. We also demonstrated that fusion of human assisted methods based on phone durations and formant features decreased equal error rate obtained by the best of these methods. We found a subset of phones with the highest speaker discriminative ability which gives a performance comparable to the case of using the full set of phones. This finding can be useful when carrying out verification on recordings of small duration that do not contain larger subsets of phones.

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

References

1. Kunzel, H., Masthoff, H., Koster, J.: The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition. Sci. Justice **35**(4), 291–295 (1995)
2. Nolan, F.: Intonation in speaker identification: an experiment on pitch alignment features. Forensic Linguist. **9**(1), 1–21 (2002)
3. 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, SPECOM 2007, Moscow, Russia, pp. 361–366 (2007)
4. Morrison, G.: Likelihood-ratio-based forensic speaker comparison using representations of vowel formant trajectories. J. Acoust. Soc. Am. **125**, 2387–2397 (2009)
5. Nolan, F., Grigoras, C.: A case for formant analysis in forensic speaker identification. J. Speech Lang. Law **12**(2), 143–173 (2005)
6. Rose, P., Osanai, T., Kinoshita, Y.: Strength of forensic speaker identification evidence: multispeaker formant-and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. Forensic Linguist. **10**(2), 179–202 (2003)
7. Becker, T., Jessen, M., Grigoras, C.: Forensic speaker verification using formant features and Gaussian mixture models. In: Proceedings of the Interspeech 2008 Incorporating SST, International Speech Communication Association, pp. 1505–1508 (2008)
8. Dellwo, V., Leemann, A., Kolly, M.-J.: Speaker idiosyncratic rhythmic features in the speech signal. In: Proceedings of Interspeech, Portland, USA, 9–13 September, pp. 1584–1587 (2012)
9. Leemann, A., Kolly, M.-J., Dellwo, V.: Speaker-individuality in suprasegmental temporal features: implications for forensic voice comparison. Forensic Sci. Int. **238**, 59–67 (2014)
10. Van Heerden, C., Barnard, E.: Speaker-specific variability of phoneme durations. S. Afr. Comput. J. (SACJ) **40**, 44–50 (2008)
11. Schwarz, P.: Phoneme recognition based on long temporal context. Ph.D. thesis, Brno University of Technology (2009)
12. Chernykh, G., Korenevsky, M., Levin, K., Ponomareva, I., Tomashenko, N.: State level control for acoustic model training. In: Ronzhin, A., Potapova, R., Delic, V. (eds.) SPECOM 2014. LNCS, vol. 8773, pp. 435–442. Springer, Heidelberg (2014)
13. Moreno, P., Joerg C., Van Thong, J.-M., Glickman, O.: A recursive algorithm for the forced alignment of very long audio segments. In: Proceedings of ICSLP 1998, Sydney, Australia, pp. 2711–2714. IEEE Press (1998)
14. Tomashenko, N.A., Khokhlov, Y.Y.: Fast algorithm for automatic alignment of speech and imperfect text data. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS, vol. 8113, pp. 146–153. Springer, Heidelberg (2013)
15. The NIST year 2010 Speaker Recognition Evaluation plan (2010). http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf