INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

Combining Gaussian mixture models and segmental feature models for speaker recognition

Milana Milošević[1], *Ulrike Glavitsch*[2]

[1]School of Electrical Engineering, University of Belgrade, Serbia
[2]Empa, Swiss Federal Laboratories for Materials Science and Technology, Switzerland
milana.milosevic@gmail.com, ulrike.glavitsch@empa.ch

Abstract

In most speaker recognition systems speech utterances are not constrained in content or language. In a text-dependent speaker recognition system lexical content of speech and language are known in advance. The goal of this paper is to show that this information can be used by a segmental features (SF) approach to improve a standard Gaussian mixture model with MFCC features (GMM-MFCC). Speech features such as mean energy, delta energy, pitch, delta pitch, the formants F1 – F4 and their bandwidths B1 – B4 and the difference between F2 and F1 are calculated on segments and are associated to phonemes and phoneme groups for each speaker. The SF and GMM-MFCC approaches are combined by multiplying the outputs of two classifiers. All the experiments are performed on the two versions of TEVOID: TEVOID16 with 16 and the upgraded TEVOID50 with 50 speakers. On TEVOID16, SF achieves 84.23%, GMM-MFCC 91.75%, and the combined approach gives 95.12% recognition rate. On TEVOID50, the SF approach gives 68.69%, while both GMM-MFCC and the combined model achieve 95.84% recognition rate.. On both databases, the number of male/female confusions decreased for the combined model. These results are promising for using segmental features to improve the recognition rate of text-dependent systems.

Index Terms: text-dependent speaker recognition, segmental features, GMM, MFCC

1. Introduction

Common approaches to automatic speaker recognition are text-independent systems where the speech utterances are unconstrained both in terms of language and lexical content [1, 2]. Few papers have been published on using the lexical content of speech for speaker recognition. Shriberg et al. extract duration, pitch and energy features for each syllable, quantize these features and form N-grams to be used in a support vector machine [3]. The problem is that an automatic speech recognition component must first compute a transcription of each utterance which is both expensive and error-prone.

In a text-dependent speaker recognition system the lexical content of the data used for system training and testing can be determined in advance. Examples are authentication systems where users speak a predefined sentence such as “My voice is my password.” to access their accounts or to enter a building. The segment boundaries have to be computed beforehand in order to derive features from the segments or syllables. For this purpose, an efficient text-to-speech alignment has to be applied. So far, there exist forced-alignment approaches to detect the phoneme boundaries in the speech utterance [4] but they are often not efficient enough. A prototype of an efficient text-to-speech alignment algorithm for such short utterances matching vowels to stable intervals [5] is under development. In this paper, we show that the speaker recognition rate is improved by using information about the lexical content of the speech utterance that is combined with classic Gaussian Mixture Models (GMM) based on MFCCs. Experiments were performed using the flexible software framework VoiceTime [6]. In previous research, we compared a segmental features approach with Gaussian Mixture Models based on MFCC features. The segmental features achieved 84.23% recognition rate, which was not so far from 91.75% obtained from GMMs. The experiments were carried out on the TEVOID database with 16 speakers [7]. For the experiments presented in this paper, we use both the 16 speaker and the 50 speaker TEVOID corpus.

2. TEVOID database

In our experiments, we used both the original TEVOID16-corpus and the extended TEVOID50 corpus [7]. TEVOID16 contains manually labeled sentences of 16 speakers. TEVOID50 has automatically aligned sentences of 50 different speakers. In both versions, there are 256 read sentences of each speaker.

Phonemes were grouped in the way they are formed as shown in Table 1.

Table 1: *Phoneme groups and phonemes*

Phoneme group		Phonemes
Vowels	open	/a/, /E/, /i/, /oe/, /o/, /u/, /y/
	closed	/a:/, /e:/, /E:/, /i:/, /o:/, /oe:/, /u:/, /y:/
Fricatives		/ch/, /f/, /sch/, /s/, /ts/
Halfvowels		/j/, /v/
Laterals		/l/
Nasals		/m/, /n/, /ng/
Vibrants		/r/
Plosives		/b/, /d/, /g/, /k/, /p/, /t/,
Diphtongs		/au/, /ei/, /eu/

3. VoiceTime Software

The VoiceTime software is a user-friendly platform for experiments on speaker recognition where users can select the recognition algorithm, training and test sets as well as the features and the combination of features [6]. For each experiment, the training and test sets are user-defined through a graphical interface.

3.1. Features

VoiceTime provides an interface for calculating the following features: MFCCs, energy, pitch, duration, formants F1–F4 and formants bandwidths B1–B4. MFCCs can be automatically calculated over the entire utterance or on particular selected phoneme groups. Energy (E) is calculated as normalized sum of squared samples of the signal. Pitch was extracted by running Praat [8] scripts through VoiceTime. The pitch sampling period was 100ms. The first four formants F1-F4 and their bandwidths B1-B4 were computed by running the Praat scripts through VoiceTime as well. For each phoneme, these eight values (F1–F4, B1–B4) are obtained from the central point of the phoneme.

3.2. Classifiers

3.2.1. Gaussian Mixture Models

A Gaussian Mixture Model [1] is considered as a reference stochastic method for our speaker identification task. A GMM speaker model consists of a finite number of mixtures of multidimensional Gaussian components. In our experiments, 50 mixtures are used.

3.2.2. Segmental feature model

The segmental features are calculated over those parts of speech that correspond to the same phoneme and to the same phoneme group (e.g. vowels, fricatives, halfvowels, etc.).

The segmental features (SF) used in the experiments are pitch, delta pitch (Δpitch), formants F1-F4, their bandwidths B1-B4, the difference between formants F1 and F2 (|F2-F1|), energy (E) and delta energy (ΔE). The same process is done for each test sequence, and then the distance measure is calculated from the given test sentence to each speaker model using the formula:

$$\Delta d(x, feature) = |mean_{model}(x, feature) - mean_{test}(x, feature)| / mean_{model}(x, feature),\ x \in \{phonemes, groups\}$$

A detailed explanation of the formula can be found in [6].

3.2.3. Hybrid approach

SF and GMM classifiers were combined by simply multiplying the outputs of these two classifiers. This is possible due to fact that both outputs are probabilities.

4. Experiments and results

We used 16 sentences for model training and 128 different sentences for testing. The total recognition rate and number of male/female confusion are given in the Table 2.

Table 2: *Speaker recognition rates using SF, GMM and Hybrid*

	TEVOID16		**TEVOID50**	
	rate	m/f conf.	rate	m/f conf.
SF	84.23%	11	68.69%	26
GMM	91.75%	6	95.84%	7
Hybrid	95.12%	0	95.84%	1

The hybrid approach gave an improvement in total recognition rate by 3.37% on TEVOID16 and 0 male/female confusions. In the case of TEVOID50, the recognition rate remained the same, but the number of male/female confusions dropped from 7 to 1. It is interesting to note that the segmental features loose on reliability with the higher number of speakers.

GMM improves on the bigger database, and this is due to the excellent discrimination of speakers added to TEVOID50. In the hybrid approach, the results are improved on the TEVOID16 database which uses manual phonetic alignment. On TEVOID50, where automatic phonetic alignment is used, only number of male/female confusion is dropped.

5. Conclusions

Speaker recognition systems can be improved by using available information such as the lexical content of speech. In this paper, we show that a segmental feature approach in combination with a classical GMM approach gives an improvement in total recognition rate for a 16 speaker database. For the extended 50 speaker database, the total recognition rate remains the same. In experiments on both databases, the number of male/female confusions drops.

In future work, more segment-based features will be defined and extracted. These features will be combined in an optimal way. In addition, an automatic text-to-speech alignment will be embedded in the software VoiceTime and algorithms for calculating pitch, formants and bandwidths will be integrated also. This way, VoiceTime will be applicable to numerous custom cases.

6. References

[1] D. A. Reynolds, and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, 1995, 3(1), pp. 72 – 83.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, 2010, 52(1), pp. 12 – 40.

[3] E. Shriberg, L. Ferrer, S. Kajabakar, A. Venkataraman, A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication*, 2005, vol. 46, pp. 455 – 472.

[4] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357 – 370, 1993.

[5] U. Glavitsch, L. He and V. Dellwo, "Stable and unstable intervals as a basic segmentation procedure of the speech signal", *in Proceedings of Interspeech*, Dresden, Germany, 2012, pp. 31 – 35.

[6] M. Milošević, U. Glavitsch, L. He, V. Dellwo, "Segmental features for automatic speaker recognition in a flexible software framework," *in 25th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA),* 2016.

[7] V. Dellwo, A. Leemann, M.-J. Kolly, "Speaker idiosyncratic rhythmic features in the speech signal," *in Proceedings of Interspeech,* Portland, Oregon, USA, 2012, pp. 1584 – 1587.

[8] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.0.28, retrieved 23 March 2017 from http://www.praat.org/