A Longest Matching Segment Approach for Text-Independent Speaker Recognition

Ayeh Jafari, Ramji Srinivasan, Danny Crookes, Ji Ming

Institute of Electronics, Communications and Information Technology
Queen's University Belfast, Belfast BT3 9DT, UK
ajafari01, r.srinivasan, d.crookes, j.ming@qub.ac.uk

Abstract

We describe a new approach for segment-based speaker recognition, given text-independent training and test data. We assume that utterances from the same speaker have more and longer matching acoustic segments, compared to utterances from different speakers. Therefore, we identify the *longest* matching segments, at each frame location, between the training and test utterances, and base recognition on the similarity of these longest matching segments. The new system scores the speaker higher who has greater number, length and similarity of matching segments. Focusing on long acoustic segments effectively exploits the spectral dynamics. We have compared our new system with the conventional frame-based GMM-UBM system for the NIST 2002 SRE task, and achieved better performance.

Index Terms: spectral dynamics, segment modeling, speaker modeling, speaker recognition

1. Introduction

The differences between speakers show up both in short-time spectra and in spectral dynamics. Many previous text-independent speaker recognition systems look at the short-time spectra of speech (e.g., the GMM-UBM [1] or SVM [2] based systems). The GMM-based systems offer smooth representations for the short-time speech spectra. However, they lack the ability to represent the dependencies of short-time spectra over long time ranges. In this paper, we study the problem of modeling long-range spectral dynamics in text-independent data for speaker recognition.

Over recent years, researchers have studied text-constrained speaker recognition within a text-independent framework. In these systems, common subword or word units, such as phones or phone trigrams, syllables, words or word *n*-grams, are detected in the training data and test data, and are used to model and recognize the speakers [3]–[5]. Alternatively, systems have been proposed that base recognition on models of acoustic segments identified either phonetically similar, or by minimum distance, obtained on the given training data set (e.g., [6], [7]). Other methods for modeling the long-range spectral dynamics of speech include the use of context-dependent phones [8] or templates [9]; the use of prosodic features such as pitch, formants and subword durations (e.g., [10]); and the characterization of speakers by using phonetic refraction or speaker-dependent pronunciation and lexical preferences, expressed as phone *n*-gram counts (e.g., [11]).

As described above, many current approaches based on equally-worded or phonetically similar segments require a separate front-end system for phone or segment identification. This increases the difficulty in developing a speaker recognition system, especially for languages without sufficient training data to build an accurate speaker-independent segment labeller. Additionally, in many current systems the front-end segment formation and the back-end speaker recognition are performed separately; there lacks a joint optimization between the two parts. In this paper, we study a different approach for segment-based speaker recognition. We aim to exploit long-range spectral dynamics without using a separate phone or segment recognizer. Given text-independent training and test data, we assume that utterances from the same speaker have more and longer matching acoustic segments, compared to utterances from different speakers. To exploit this, we use an approach based on the maximum posterior probability criterion. We compare a test utterance directly with the training utterances to find their *longest* matching acoustic segments at each frame location. Then we perform recognition based on the similarity of the longest matching segments found. Therefore our system scores the speaker higher who has greater number, length and similarity of matching segments. The new approach can be viewed as an extension of the phonetic refraction approach (e.g., [11]) to acoustic segments. Distinct acoustic events are likely to occur in a given speaker's utterances which would indicate that different speakers may have different acoustic preferences. Here an acoustic event can be an arbitrary-length segment of consecutive frames, which may be any sound made by a speaker, not limited to a subword unit, and not necessarily phonetically transcribable. Focusing on the longest matching segments best exploits the spectral dynamics. For convenience, we call our new approach the longest matching segment (LMS) approach.

2. The longest matching segment approach

The new LMS approach includes three components: 1) a novel speaker model combining statistical and template-based approaches to capture up to sentence-long spectral dynamics in the training data, 2) an algorithm for identifying the longest matching segments between the training and test sentences, and 3) a method of forming the recognition scores based on the longest matching segments found. The following gives details of each of these components.

2.1. A novel example-based speaker model

We model the *complete* spectral dynamics in each training sentence, such that any segment of any length in the sentence, up to the complete sentence, can be used as a whole unit to identify the corresponding units/segments in the test speech. We use a new example-based approach, as opposed to conventional templates, to build the models. The new approach includes

This work was supported by the UK EPSRC grant EP/G001960.

two steps. First, for each speaker, we train a GMM using all the speaker's training sentences, as in traditional GMM-based recognition systems. Second, based on the GMM, we build a further model for each training sentence to represent the complete spectral dynamics of the sentence. Let G_λ represent the GMM for speaker λ, modeling the probability distribution of short-time speech spectral vectors x for the speaker:

$$G_\lambda = \{g(x|k,\lambda), w(k|\lambda) : k = 1, 2, ..., K\} \qquad (1)$$

where $g(x|k,\lambda)$ is the k'th Gaussian component and $w(k|\lambda)$ is the corresponding weight. Let $\mathbf{x} = \{x_i : i = 1, 2, ..., I_\mathbf{x}\}$ represent a training sentence from speaker λ with $I_\mathbf{x}$ frames and x_i being the frame at time i. Based on G_λ, we can obtain a new representation for $\mathbf{x}$, by taking each frame from $\mathbf{x}$ and finding the Gaussian component in G_λ that produces maximum likelihood for the frame. This results in a time sequence of maximum-likelihood Gaussian components $g(x|k_{\mathbf{x},i},\lambda)$, $i = 1, 2, ..., I_\mathbf{x}$, where $g(x|k_{\mathbf{x},i},\lambda)$ is the Gaussian component in G_λ that produces maximum likelihood for the ith frame x_i in training sentence $\mathbf{x}$. We will use this maximum-likelihood Gaussian sequence as a model for training sentence $\mathbf{x}$. In the model, the individual Gaussian components represent the probability distributions of the short-time speech spectra that form this sentence, and the time sequence of the Gaussian components captures the full spectral dynamics, from acoustic to lexical and to language, that join together the appropriate short-time spectra to form the specific sentence. This sentence model, i.e., the maximum-likelihood Gaussian sequence, can be represented by the corresponding time sequence of indices $(\mathbf{k}_\mathbf{x}, \lambda)$, which is defined as:

$$(\mathbf{k}_\mathbf{x}, \lambda) = \{(k_{\mathbf{x},i}, \lambda) : i = 1, 2, ..., I_\mathbf{x}\} \qquad (2)$$

where $(k_{\mathbf{x},i}, \lambda)$ indexes a Gaussian $g(x|k_{\mathbf{x},i},\lambda)$ in G_λ, that gives maximum likelihood for frame x_i in sentence $\mathbf{x}$.

Given a test sentence, traditional GMM-based systems perform recognition based on G_λ. Our new approach performs recognition through the *sentence* models $(\mathbf{k}_\mathbf{x}, \lambda)$, for all the training sentences $\mathbf{x}$ for each speaker λ. The difference is important: the GMM approach allows consecutive frames in a test sentence to be matched by any sequences of Gaussian components from G_λ, while the new approach forces the match to the Gaussian sequences forming the training sentences $\mathbf{x}$. The new approach, thus, exploits similarities both between the short-time spectra and between their dynamics. In the following, we will describe an algorithm for optimizing the discrimination by forcing the comparison between long acoustic segments between the training and test sentences, based on the training sentence models (2).

In recent years, there have been studies in using templates as an alternative to GMM for speaker recognition (e.g., [9]). While templates can capture long-range spectral dynamics, they lack smoothness (and hence robustness) in representing the short-time speech spectra, which are subject to random variations. The above sentence model, (1) and (2), combines the advantages of GMM and templates. It offers both a smooth representation for the short-time spectra and a sentence-long representation of the spectral dynamics.

2.2. Identifying the longest matching segments

Given a test sentence $\mathbf{y} = \{y_t : t = 1, 2, ..., T\}$, of T frames with y_t being the frame at time t, we will compare it to the training sentences of each speaker for recognition. Since a segment of consecutive speech frames, when treated as a whole unit, provides greater speaker discrimination than the individual frames, we seek the *longest* matching segments between the training and test sentences to compare, as a means of maximizing the discrimination given text-independent data. The training sentence model (2) is used to formulate the comparison.

Let $\mathbf{y}_{t:\tau} = \{y_\epsilon : \epsilon = t, t+1, ..., \tau\}$ represent a test segment taken from test sentence $\mathbf{y}$ and consisting of consecutive frames from time t to τ. Let $(\mathbf{k}_{\mathbf{x},u:v}, \lambda) = \{(k_{\mathbf{x},i}, \lambda) : i = u, u+1, ..., v\}$ represent a training segment taken from training sentence model $(\mathbf{k}_\mathbf{x}, \lambda)$ and modeling consecutive frames from u to v in training sentence $\mathbf{x}$ from speaker λ. We measure the similarity between the two segments by calculating the posterior probability $P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau})$. Assuming an equal prior probability for all the training segments, this can be written as:

$$P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau}) = \frac{p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda)}{p(\mathbf{y}_{t:\tau})}$$
$$= \frac{p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda)}{\sum_{\lambda'}\sum_{\mathbf{x}'}\sum_{u',v'} p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x}',u':v'}, \lambda') + p(\mathbf{y}_{t:\tau}|\phi)} \qquad (3)$$

In (3), $p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda)$ is the likelihood function that the test segment $\mathbf{y}_{t:\tau}$ matches the training segment $(\mathbf{k}_{\mathbf{x},u:v}, \lambda)$. Assuming independence between the frames within a segment, this likelihood function can be calculated using the Viterbi algorithm and can be expressed as:

$$p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda) = \prod_{\epsilon=t}^{\tau} g(y_\epsilon|k_{\mathbf{x},i_\epsilon}, \lambda) \qquad (4)$$

where i_ϵ is the most-likely time map function between the test frames y_ϵ and the training frames modeled by $(k_{\mathbf{x},i_\epsilon}, \lambda)$, assuming that $i_t = u$ and $i_\tau = v$. If we view the segmental spectral dynamics, associated with the training segment $(\mathbf{k}_{\mathbf{x},u:v}, \lambda)$ in this example, as "text" dependence, then (4) gives a "text-dependent" likelihood of the test segment. In the denominator of (3), the first term includes all the training segments, from all the training sentences of all the speakers with all possible segment locations and lengths, that are likely to match the test segment $\mathbf{y}_{t:\tau}$. The second term $p(\mathbf{y}_{t:\tau}|\phi)$ represents the likelihood that $\mathbf{y}_{t:\tau}$, *as a whole unit*, is not seen in any of the speakers' training sentences (assuming an equal prior P). This likelihood of unseen segments can be suitably modeled by using a "text-independent" model, for example, a GMM-based UBM trained with data from all the speakers.

Assume that $\mathbf{y}_{t:\tau}$ and $(\mathbf{k}_{\mathbf{x},u:v}, \lambda)$ are two matching segments in the sense that the test segment $\mathbf{y}_{t:\tau}$ achieves the highest likelihood $p(\mathbf{y}_{t:\tau}|\mathbf{k}_{\mathbf{x},u:v}, \lambda)$ compared to all the other training segments, including ϕ. Then it can be shown that [13]

$$P(\mathbf{k}_{\mathbf{x},u:i_\epsilon}, \lambda|\mathbf{y}_{t:\epsilon}) \leq P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau}) \quad \text{for } \epsilon \leq \tau \qquad (5)$$

In other words, the posterior probability increases when longer segments are matched. Thus, we can use the maximum values of the posterior probability to locate the longest matching segments between the test sentence and the training sentences, to be used for recognition. The following describes the algorithm.

2.3. Recognition based on the longest matching segments

Consider verifying the test sentence $\mathbf{y} = \{y_t : t = 1, 2, ..., T\}$ against speaker λ. At each time t, we can find a longest test segment $\mathbf{y}_{t:\tau_{\max}}$ from t and the corresponding matching training segment $\mathbf{k}^t_{\mathbf{x},u:v}$ from λ, by maximizing the posterior probability. This can be expressed as:

$$P(\mathbf{k}^t_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau_{\max}}) = \max_{\tau} \max_{\mathbf{k}_{\mathbf{x},u:v}\in\lambda} P(\mathbf{k}_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau}) \quad (6)$$

That is, $\mathbf{k}^t_{\mathbf{x},u:v}$ is obtained by finding a most-probable training segment for each fixed-length test segment $\mathbf{y}_{t:\tau}$, and then finding the maximum test-segment length (i.e., $\tau_{\max}$) resulting in the maximum posterior probability, from all the training sentences for speaker λ. The posterior probability $P(\mathbf{k}^t_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau_{\max}})$ gives the similarity of two longest matching segments found between the training and test sentences, in terms of maximum posterior probability (or similarity) both between their short-time spectra and between their spectral dynamics. We will use $P(\mathbf{k}^t_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau_{\max}})$ at each time t to form a sentence score for verification, assuming that utterances from the same speaker exhibit a higher count of long matching segments (and hence a higher posterior probability) than utterances from different speakers. Let $\Gamma(\lambda;\mathbf{y})$ be the score for speaker λ given the test sentence $\mathbf{y}$. This is obtained by summing the logarithmic posterior probabilities of the longest matching segments corresponding to each of the frames in the test sentence:

$$\Gamma(\lambda;\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \log P(\mathbf{k}^t_{\mathbf{x},u:v}, \lambda|\mathbf{y}_{t:\tau_{\max}}) \qquad (7)$$

As shown in (7), each test frame is scored through a longest matching segment. This helps to capture the long-range spectral dynamics about the frame available in the training data.

3. Model adaptation

We can use an adaptation method to derive each speaker's GMM G_λ, (1), given limited training data. The method is similar to the one used in the conventional GMM-UBM systems [1]. First, a UBM is estimated by using all the speakers' training data. Then, the speaker λ's GMM, G_λ, is obtained by taking the UBM as an initial model and updating its parameters using the given training sentences for the speaker, with the EM algorithm. Finally, we obtain a sentence model, (2), for each training sentence by identifying the maximum-likelihood Gaussian from the adapted G_λ for each frame, to be used as the model of the speaker for recognition. In the new algorithm the mean and covariance of the adapted model G_λ are obtained by interpolating the EM-algorithm based mean and covariance with the UBM mean and covariance respectively.

4. Experimental studies

Experiments were conducted on the NIST 2002 SRE database for the one speaker detection task. The task contains cellular conversational speech data from 330 speakers (139 male, 191 female). Each speaker contributed about two minutes of data for training; there were a total of 3570 sentences (1442 male, 2128 female) with variable durations from 15 to 45 seconds for testing. The silence-removed speech was divided into frames of 20 ms with a frame period of 10 ms. Each frame was modeled by using a 26-element feature vector, consisting of 13 MFCC (C_0–C_{12}) and their first-order derivatives. Cepstral mean subtraction was applied to each sentence.

We conducted the experiments for two types of speaker models. The first type of model contained 128 mixtures for each speaker and was trained directly from the training data for the speaker. The second type of model contained 1024 mixtures for each speaker and was obtained by adapting from a gender-dependent UBM trained on this database.

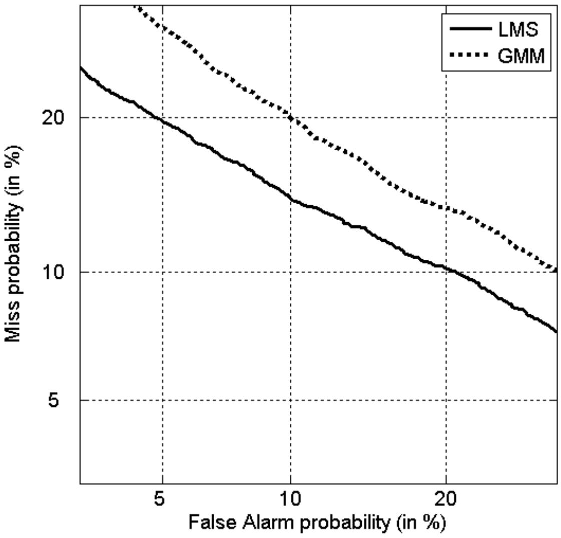

Figure 1: *DET curves for non-adapted speaker models, comparing the new LMS approach with a baseline GMM.*

Table 1: *EER (%) for non-adapted speaker models, for the new LMS approach and a baseline GMM.*

System	EER	Relative improvement
GMM	15.50	-
LMS	12.83	17.2%

Table 2: Examining the LMS performance for adapted speaker models when adapting different parameters.

Parameters adapted in the LMS system	EER %
mean + covariance	9.36
mean only	10.50

4.1. Results for non-adapted speaker models

First we compare the new LMS system with the baseline GMM system with each speaker's model trained independently. Fig. 1 shows the DET curves, and Table 1 presents the corresponding equal error rates (EER). The new LMS system showed a clear improvement over the baseline GMM system, reducing the EER by over 17%.

4.2. Results for adapted speaker models

Then, we compare the new LMS system with the baseline GMM-UBM system, with the speaker models adapted from the UBM. The baseline system was adapted using the algorithm described in [1], and produced the best results with adapting the mean vectors only. Our new system was adapted using the algorithm described in Section 3. We have tested modifying different parameters, particularly, mean and covariance, for the LMS system in the adaptation and the results are shown in Table 2. As can be seen, the LMS method for adapted speaker models achieved the best overall performance for adapting both mean and covariance. This is in contrast with the baseline GMM-UBM system. Fig. 2 shows the DET curves, and Table 3 presents the corresponding EER comparison. The results

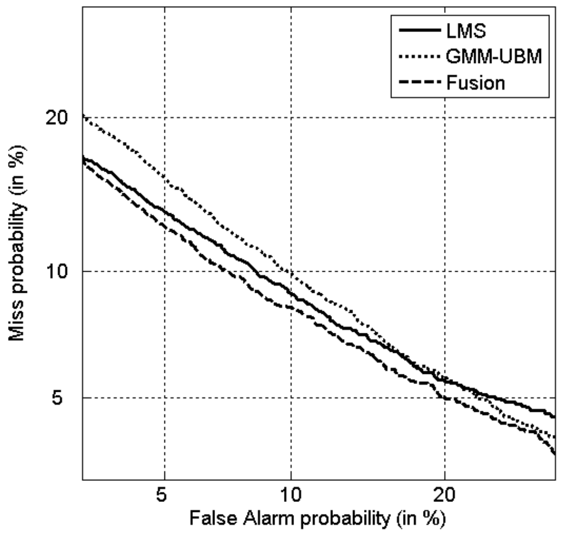

Figure 2: *DET curves for adapted speaker models, comparing the new LMS approach, a baseline GMM-UBM, and their fusion.*

Table 3: *EER (%) for adapted speaker models, for the new LMS system, a baseline GMM-UBM, and a fused system.*

System	EER	Relative improvement
GMM-UBM	9.95	-
LMS	9.36	5.9%
Fusion	8.73	12.2%

we obtained for the baseline GMM-UBM system were consistent with those reported in [12]. Again, the new LMS system outperformed the GMM-UBM baseline. In fact, to the authors' knowledge, it is one of the few stand-alone systems that could outperform the GMM-UBM baseline by modeling text dependency, in the recent NIST SRE tasks.

The LMS approach performed recognition based on the longest matching segments between the test sentence and the training sentences of each hypothesized speaker. Fig. 3 shows the histograms of the lengths of the longest matching segments, measured in number of frames, found by the algorithm for the true speakers and imposters. A significant observation can be drawn from the histograms. That is, more longer matching segments were found between the speeches for the true speakers than for the imposters. For example, there were about 66% of the matching segments for the true speakers that were two or more frames long, while for imposters there were only about 48%. This confirms our intuition that utterances from the same speaker should have higher counts of similar long-span spectral dynamics than utterances from different speakers.

Finally, we combined the new LMS system, which emphasizes the similarity of long-range spectral dynamics, with the GMM-UBM system, which emphasizes the similarity of short-time spectra, by linearly fusing their sentence-level scores. The results are shown in Fig. 2 and Table 3. The fused system further reduced the ERR by over 6% compared to the LMS system, and resulted in an overall 12.2% relative improvement compared to the baseline GMM-UBM.

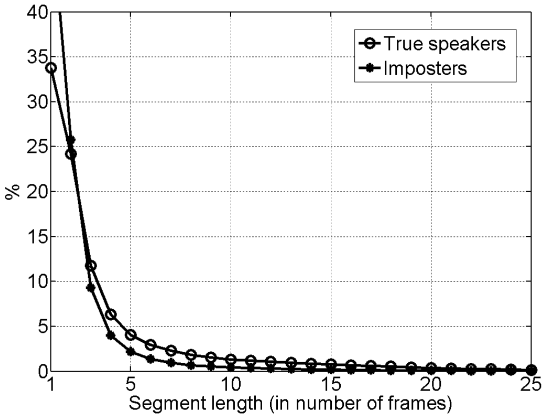

Figure 3: *Histogram of the length of the longest matching segments between the training and test sentences, found by the proposed LMS algorithm.*

5. Conclusions

We described an algorithm capable of detecting the longest matching segments between training and test utterances, and proposed a speaker recognition system based on the longest matching segments found. The new algorithm aims to more effectively capture the differences in long-range spectral dynamics between different speakers, given text-independent speech data. Experimental results on the NIST SRE 2002 database indicate that the new system outperformed the baseline systems. Further improvement was obtained by fusing with the baseline.

6. References

[1] Reynolds, D. A., *et al.*, "Speaker verification using adapted Gaussian mixture models", Digital Signal Process., 10:19-41, 2000.

[2] Campbell, W. M., *et al.*, "Support vector machines for speaker and language recognition", Computer and Speech Language, 20:210-229, 2006.

[3] Sturim, D. E., *et al.*, "Speaker verification using text-constrained Gaussian mixture models", ICASSP,'2002, 667-680.

[4] Aronowitz, H., *et al.*, "Text independent speaker recognition using speaker dependent word spotting", ICSLP'2004.

[5] Bocklet, T., Shriberg, E., "Speaker recognition using syllable-based constrants for cepstral frame selection", ICASSP'2009, 4525-4528.

[6] Gerber, M., *et al.*, "Fast search for common segments in speech signals for speaker verification", Interspeech'2008, 375-378.

[7] Tsao, Y., *et al.*, "An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition", ICASSP'2010, 4422-4425.

[8] Stolcke, A., *et al.*, "MLLR transforms as features in speaker recognition", Interspeech'2005.

[9] Gillick, D., Stafford, S., and Peskin, B., "Speaker detection without models", ICASSP'2005, 757-760.

[10] Adami, A., Mihaescu, R., Reynolds, D., and Godfrey, J., "Modeling prosodic dynamics for speaker recognition", ICASSP'2003.

[11] Andrews, W. D., *et al.*, "Gender-dependent phonetic refraction for speaker recognition", ICASSP'2002, 149-152.

[12] Barras, C., Gauvain, J., "Feature and score normalization for speaker verification of cellular data", ICASSP'2003, 49-52.

[13] Ming, J., "Maximizing the continuity in segmentation - a new approach to model, segment and recognize speech", ICASSP'2009, 3849-3852.

.