DISCRIMINATIVE POWER OF TRANSIENT FRAMES IN SPEAKER RECOGNITION

Jérôme Louradour, Khalid Daoudi, Régine André-Obrecht

IRIT-Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse - France

ABSTRACT

In speaker recognition, several recent studies attempt to integrate prior knowledge in order to make better distinction between speakers. This paper study the relative speaker discriminative power of speech transient and steady zones. An automatic segmentation is used to localize these two types of zones. Experiments are carried out with NIST 2003 speaker evaluation database. They show that transient frames, in the neighborhood of segment boundaries, are more speaker discriminative than the middle frames of long segments which correspond to the steady parts of phones. In addition this study leads to higher efficiency system (than the baseline one) by processing roughly three times less data during the scoring stage.

1. INTRODUCTION

In speaker verification, it is crucial to properly select zones of the signal that will be processed for model training, as well as for the scoring of a test utterance. A first requirement is to detect speech zones by removing periods of silence and noise. It has already been observed that the choice of the speech activity detector affects considerably the overall performance of a speaker recognition verification system [1]. Improvement can also be achieved by discarding out-lier observations in scoring. For instance, approaches known as score pruning [2, 3] examine how to reject abnormal frame scores in order to make the system more robust. Another recent trend consists of attaching more or less importance to zones of speech observations without necessarily rejecting them.The allocated relative importance depends on the presumed speaker separating power of an observation. Two alternatives are conceivable. The first one attempts to make models more accurate during enrollment. For instance, [4] examines how to improve performance by adjusting the weights of the GMM according to the corresponding phoneme class. A second alternative consists of weighting the influence of the different observations in the decision score, during testing. For instance, [5] and [6] suggest to compute a weighted mean of the frame scores instead of the classical average procedure. In these studies, weights computation is based on prior knowledge about respectively target speaker pitch distribution and score statistics.

In the same spirit, we proposed in [7] a weighting procedure where each frame is allocated a weight depending on its distance to phonetic targets. Motivated by this work, the main scope of the present paper is to show, through an empirical study, that speaker discriminative information actually lies in transient zones and that steady ones are less sensitive to speaker inter-variability. Determination of these zones is performed using an automatic segmentation procedure. From a phonetic point of view, this means that essential discriminative information lies in the way speakers move from one sound to the other, rather than in the phonetic targets themselves. We note here that the opposite is true when the task is speech recognition (in [8] it is shown focusing on steady zones with a segmental approach yields comparable performance to the centisecond approach). From a computational point of view, this means that steady zones can be discarded in test without a loss in performance, thus leading to higher efficiency.

The paper is organized as follows. In section 2, we recall the speech segmentation procedure. Section 3 describes the experimental set up. In section 4 we present several experiments which show the discriminative power of transient zones with respect to steady ones.

2. A PRIORI SPEECH SEGMENTATION

In order to localize transient zones, we use an a priori segmentation based on the Kullback divergence, which is explained in [9]. This segmentation has the advantage of processing the signal without relying on the extraction of explicit features. The underlying theory only assumes that the speech signal is a sequence of quasi stationary segments: the output segments are expected to represent sub-phonetic units (Fig.1).

Whereas the output information is more basic than a phonetic classification, the resulting segmentation technique is quite robust to transmission and recording conditions. Another advantage of this technique is to give a good accuracy on the temporal location of the boundaries of the segments. In fact, two kinds of segments appear:

- long segments (often longer than 40 ms) which correspond to the steady part of phones,
- small segments (about 20 ms) which are transitory zones or which reveal the superposition of articulator movements.

Transient zones stands inside small segments and in the neighborhood of segment boundaries. In the following, we call a ”transient frame” a frame located in a small segment or around a segment boundary. A “steady zone” is obviously all frames except transient ones.

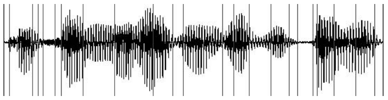

Fig. 1. View of a speech signal segmentation

3. EXPERIMENTAL SETUP

This section presents our baseline system as well as the evaluation corpus.

3.1. Baseline System

The baseline system, known as UBM-GMM system ([10]), implements, for each feature vector from a test sequence, a likelihood ratio test with a model representing the claimed speaker, obtained from a previous enrollment, and a Universal Background Model representing potential impostors. Feature vectors are extracted on regularly step frames. In the reference approach, the score for a sequence of such observations $O = O_1, ..., O_T$ is computed as the average frame log-likelihood ratio :

$$S(O) = \frac{1}{T}\sum_{t=1}^{T} s_t = \frac{1}{T}\sum_{t=1}^{T}(p(O_t/\Lambda_{tar}) - p(O_t/\Lambda_{UBM})) \quad (1)$$

where T is the length of the sequence after silence removal, O_t is a feature vector at frame t, s_t is the corresponding frame score, and Λ_{tar} (resp. Λ_{UBM}) are parameters of the target model (resp. world model).

The techniques we use for the three main components, front-end processing, modeling, and scoring, are briefly described next.

3.1.1. Front-end Processing

First, the speech is pre-emphasized, and 12 MFCC, are extracted on 16-ms Hamming window, processing at a 10-ms frame rate. The 12 derivatives and the energy logarithm derivative are also added. Then, speech activity detector is executed by discarding segments S_i in which signal standard deviation σ_{S_i} is lower than $\alpha \min_{S_i}(\sigma_{S_i})$. α is a scale factor that is set at 2.5 in our experiments.

Finally, picked out 25-dimensional feature vectors are warped ([11]) over 300 frame windows. Moreover, while processing test utterances, we also extract from the segmentation, and store for each frame its position within the containing segment.

3.1.2. Modeling and Scoring

The parameters of gender-dependent background model is estimated using 4 hours of speech from Switchboard-I corpus, including 3 hours of cellular data. The initialization of the 512 diagonal-covariance components GMM is done by a Vector Quantization algorithm, and EM iterations were then performed. Parameters of each target GMM were derived from the UBM by adapting only mean vectors with a MAP criterion.

During testing, only the 10-best scoring components from the UBM are used to calculate the frame log-likelihood ratio.

3.2. Evaluation Corpus

Experiments are conducted on a speaker verification task using the male limited data setup of the 2003 NIST Speaker Recognition Evaluation. About two minutes of speech is available to learn each of the 149 target speaker models, originating from landline or cellular telephone handset. For the testing, there are 1213 target trials and 13413 impostor trials, including both matched-handset and mismatched-handset conditions. Trials lengths vary from three seconds to one minute.

4. ASSESSMENT OF THE DISCRIMINATIVE POWER OF TRANSIENT FRAMES

In this section, experiments are performed to study the discriminative power of transient frames. All systems, derived from the baseline, are identically trained and only differ in the scoring procedure.

4.1. Weighted Scoring

In [7] we introduce a weighting procedure that allocates weights to the frame scores in function of their corresponding position within the segmentation. Weighting is performed by replacing the baseline scoring (Eq. 1) by:

$$S(O) = \frac{1}{\sum_{t=1}^{T}\omega_t}\sum_{t=1}^{T}\omega_t.s_t \quad (2)$$

where ω_t are positive reals. We observe that allocating higher weights to frames localized near segment boundaries slightly improve the speaker discriminant power, while allocating higher weights to steady zones significantly degrades performances, as it is shown in Fig.2. For this experiments we use :

$$\omega_t^+ = max(\frac{1}{4}, 1 - \frac{3}{16}d_t) \quad (3)$$

$$\omega_t^- = min(1, \frac{1}{4} + \frac{3}{16}d_t) \quad (4)$$

where d_t is the distance between frame t center and the nearest segment boundary, at the frame scale.

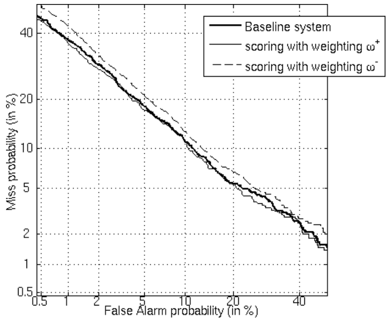

Fig. 2. DET plots with a weighted scoring

Motivated by this observation, we decided to (empirically) evaluate the discriminative information contained in segment boundaries and steady zones, respectively. This is the scope of the next subsection.

4.2. Scoring with one feature vector per segment

We conceive several systems where the scoring stage retains only a sub-sampling of the set of input test feature vectors. This sub-sampling is done by synchronizing the regular step frame extraction with the result of the segmentation: feature vectors are picked out regarding to the position of the corresponding frame center within the containing segment. We considere four fast-scoring systems for which the sub-sampling process is respectively based on the following heuristics :

(S1) Retaining only the first frame extracted from each segment marked as speech.

(S2) Retaining only the last frame for each segment.

(S3) Retaining only the closest frame to segment middle.

(S4) Picking out randomly the same number of frames as the number of segments.

Note that, given a test utterance, all these systems compute a score from the same number of frame scores. On average, this is estimated to represent 17% (i.e. about one out of six) of the original set used in the baseline (B). Results are shown in Fig.3. Whereas system (S3) gives more or less the same results as the random system (S4), one can see that (S1) and (S2) perform better that (S3). However, they do not reach the same performance as the baseline system (B).

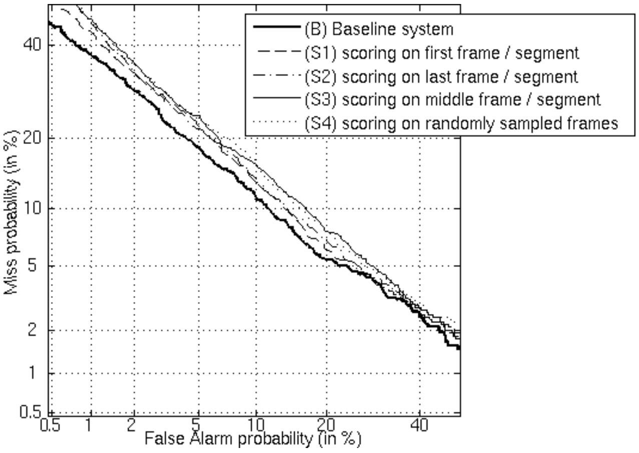

Fig. 3. DET plots using a subsampling of frame scores

The loss in performance (w.r.t. (B)) was predictible since we process six time less data during the scoring stage. But the fact that (S1) and (S2) outperform (S3) shows that each frame does not bring the same relative amount of information for speaker discrimination. Actually, transient frames (i.e. the ones that precede or follow a segment boundary) are more speaker discriminative than steady ones (near segment middles). The scope of the next subsection is to evaluate the discriminative information contained in these transient frames.

4.3. Scoring on transient frames

We compare the baseline system (B) with a system (S5) that computes a score from all transient frames located around segment boudaries. That is, all feature vectors collected by bringing together (S1) and (S3) are scored in (S5). We thus retain two frames per segment (first and last frames) and only one for very short segments. On average, this amounts to estimate a score on 32% of the original frame population of the baseline (B).

Fig.4 depicts the results, from which one can see that scoring only on transient frames leads roughly to the same performance as when taking every frames into account. Another experiment have shown that adding middle frame scores (from S2) to the amount of scoring data collected in (S5) does not improve performance (this experiment is not showed as DET curves are quite confounded). These results suggest that transient frames contain all speaker discriminative information, and that steady zones do not bring further information and can thus be disgarded in the scoring stage.

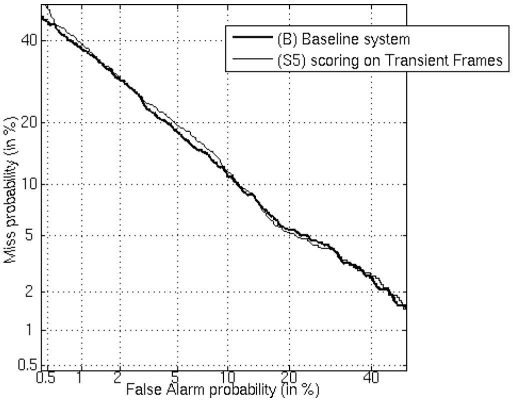

Fig. 4. DET plots when scoring on transient frames only

An immediate consequence of this result is the computational gain. Indeed, we achieve almost the same performance as the baseline system by processing 32% of the original frame scores. From a phonetic point of view, this result confirms our initial intuition that the essential speaker discrimative information lies in the transitions between phonetic targets, i.e., in the way a speaker move from one sound to another. We believe that this is an important result that can be further exploited in speaker verification, identification, and tracking tasks.

4.4. T-normalization of transient frames scores

T-norm is a score normalization process that improves significantly performance in the low False Alarm rate region [12]). It is widely used in speaker verification and works as follows. During test, a set of example impostor models is used to calculate impostor scores for a given test utterance. Mean μ_S and variance σ_S are then estimated from these scores and are used to perform the score normalization.

Since T-norm amounts to compute many scores for a given test utterance, it is quite computationally expensive. Actually, during test this step is the one that requires most computation time. From this viewpoint, sub-sampling the set of feature vectors to be scored as suggested in the previous sub-section is worthwhile when applying such a score normalization.

Fig.5 displays the DET curves obtained when applying T-norm to the Baseline (boldline) and to (S5) (dashed line). One observes that applying a T-norm procedure does not lead to the same performance of the baseline system with T-norm (bold line). This discrepancy in the results do not really raise questions about the discriminative power of transient frames, but can be explained by statistics as follows. In (S5^T), the parameters (μ_S,σ_S) to be estimated for the T-norm are subject to higher variability because of the lower number of frame scores to be averaged. Indeed, let's consider a frame score X as a random variable with mean μ_X and variance σ_X. The sampling distribution of the sequence score on an impostor model, $S = \frac{X_1+...+X_n}{n}$ (n being the length of the sequence), has a mean of μ_X and a standard deviation of $\frac{\sigma_X}{\sqrt{n}}$, which is called the standard error of S. In the same way, considering N impostor models, one can estimate the standard error of the μ_S (resp. σ_S) as being $\frac{\sigma_X}{\sqrt{n.N}}$ (resp. $\frac{\sigma_X}{\sqrt{2n.N}}$).

As the system (S5^T) works with about three times less frame scores than the baseline system, we can see that the standard error for the statistics required for T-norm will be $\sqrt{3}$ times more elevated. Thus, in order to make a fair comparison, we compare (S5^T) to a baseline system with a T-norm estimated on three times less impostor models (solid line). By doing so, the two systems not only have the same theorical standard error but also the same computation complexity. One can see that on this fair comparison, (S5^T) outperforms the baseline. This suggests that it is better to score on transient frames while exploiting a wider range of impostor speakers than the reverse.

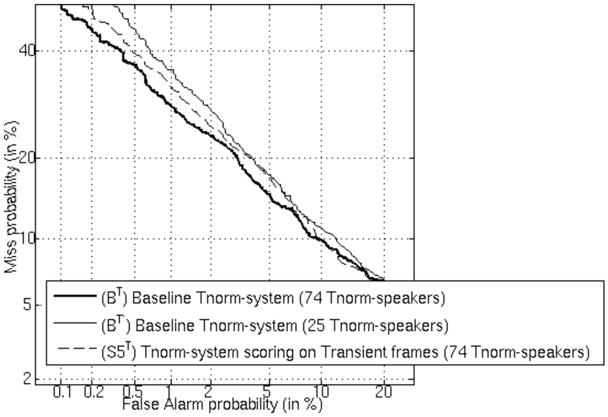

Fig. 5. DET plots showing effect of T-norm when scoring on transient frames

5. CONCLUSION

One of the main challenges is speaker recognition (and all classification tasks) is to infer parsimonious and discriminative information from the available data. This is indeed crucial for classification accuracy, robustness and efficiency. In this work, we first used an automatic speech segmentation procedure to identify transient and steady zones of speech signals. We then carried out an empirical study to show that the essential information to distinguish between speakers is contained in transient zones. As a consequence, a significant computation time was saved by discarding steady zones in the scoring stage, without a loss in performances.

At the time of writing, we are conducting further analysis to confirm this result (using SVM classifiers, by focusing on transient zones in training,...). The results of this analysis will be the purpose of future communications.

6. REFERENCES

[1] R.D. Zilca, J.W. Pelecanos, Chaudhari U.V., and Ramaswamy G.N., "Real Time Robust Speech Detection for Text Independent Speaker Recognition," in *Proc. Odyssey-04*, 2004.

[2] L. Besacier and J.-F. Bonastre, "Frame pruning for speaker recognition," in *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, 1998.

[3] K. Chen, "Towards better making a decision in speaker verification," *Pattern Recognition*, , no. 36, 2003.

[4] R. Faltlhauser and G. Ruske, "Improving speaker recognition performance using phonetically structured gaussian mixture models," in *Proc. Eurospeech*, 2001.

[5] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "On the Use of Quality Measures for Text-Independent Speaker Recognition," in *Proc. Odyssey-04*, 2004.

[6] M. Mak, M. Cheung, and S. Kung, "Robust speaker verification from gsm-transcoded speech based on decision fusion and feature transformation," in *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, 2003.

[7] J. Louradour, R. Andre-Obrecht, and Daoudi K., "Segmentation and relevance measure for speaker verification," in *Proc. ICSLP 2004*, to appear.

[8] V. Le Maire, R. Andre-Obrecht, and D. Jouvet, "An acoustic-phonetic decoder based on an automatic segmentation algorithm," in *Eurospeech I*, 1989.

[9] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Speech and Audio Proc.*, vol. 36, no. 1, 1988.

[10] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[11] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey 2001*, 2001.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.