FRAME PRUNING FOR SPEAKER RECOGNITION

L. Besacier, J.F. Bonastre

LIA/CERI
339, chemin des Meinajaries BP 1228
84911 Avignon Cedex 9 (France)
(laurent.besacier,jean-francois.bonastre)@univ-avignon.fr

ABSTRACT

In this paper, we propose a frame selection procedure for text-independent speaker identification. Instead of averaging the frame likelihoods along the whole test utterance, some of these are rejected (pruning) and the final score is computed with a limited number of frames. This pruning stage requires a prior frame level likelihood normalization in order to make comparison between frames meaningful. This normalization procedure alone leads to a significative performance enhancement. As far as pruning is concerned, the optimal number of frames pruned is learned on a tuning data set for normal and telephone speech. Validation of the pruning procedure on 567 speakers leads to a 27% identification rate improvement on TIMIT, and to 17% on NTIMIT.

1. INTRODUCTION

Most of speaker identification systems use an averaging of the frame scores to compute a global score and to take a decision with regard to the whole test utterance. This stage, which can be called *accumulation* is an arithmetic mean in the majority of cases. However, there are several ways to cope with the accumulation problem: normalizing the frame scores [8], replacing the score for a frame with a measure of confidence that the frame was spoken by the target speaker [6]. In this work, we investigate accumulation using a hard threshold approach since some frame scores are knocked out from consideration (pruning) and the final decision is taken with a subset of these scores.

Our motivation in the use of this system is to automatically extract from the input speech signal the part that at best contributes to identify a speaker. We have already investigated the selection of the most speaker specific frequency segments (subbands) for speaker identification [2]. This work is the counterpart of the previous one in the time domain.

This method should be robust in the case of noise occurring in a given time period since the least reliable frames can be removed. Even in the case of clean speech, some frames of a speaker test utterance can be simply more similar to another speaker model than to the right speaker model itself. Removing these error-prone frames should lead to a more robust decision.

In *Section 2*, we propose a formalism to describe our segment-based speaker recognition system in which a segment can be a frame or a group of frames. In *Section 3*, we describe the pruning stage which requires a previous normalization of the scores. The normalized scores proposed are interpreted as likelihood ratios. Experiments intended to find the optimal size of segments and the optimal number of frames kept are described in *Section 4*. The hyper parameters (size of segments and number of frames) which lead to the best performances are validated on TIMIT and NTIMIT databases (*Section 5*). Finally, we summarize our main results and outline the potential advantages of the pruning procedure in *Section 6*.

2. FORMALISM

The gaussian modeling is more precisely described in [3] and [6]. Let $\{x_t\}_{1\le t\le M}$ be a sequence of M vectors resulting from the *p*-dimensional acoustic analysis of a speech signal uttered by speaker X. These vectors are summarized by the mean vector $\bar{x}$ and the covariance matrix X:

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \quad \text{et} \quad X = \frac{1}{M}\sum_{t=1}^{M}(x_t-\bar{x})(x_t-\bar{x})^T \qquad (1)$$

Similarly, for a speech signal uttered by speaker Y, a sequence of N vectors $\{y_t\}_{1\le t\le N}$ can be extracted.

Supposing that all acoustic vectors extracted from the speech signal uttered by speaker X are distributed like a Gaussian function, the likelihood of a single vector y_t uttered by speaker Y is:

$$G(y_t / X) = \frac{1}{(2\pi)^{p/2}(\det X)^{1/2}} e^{-\frac{1}{2}(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \qquad (2)$$

If we assume that all vectors y_t are independent observations, the average log-likelihood of $\{y_k\}_{t+1\le k\le t+T}$ on a segment of T frames can be written:

$$\overline{G_X}(y_{t+1}^{t+T}) = \frac{1}{T}\log G(y_{t+1}\ldots y_{t+T}/X) = \frac{1}{T}\sum_{k=1}^{T}\log G(y_{t+k}/X) \quad (3)$$

The log-likelihoods of all segments are then accumulated over the whole test utterance of N frames, to form a final score for each speaker model :

$$\tilde{G}_X(y_1^N) = \underset{t\in\{0\ldots n-1\}}{ACC}\left\{\overline{G}_X(y_{tT+1}^{tT+T})\right\} \quad (4)$$

where *ACC* is the accumulation function, T is the number of frames in a segment and *n* is the number of segments; the total number of frames is then *N=nT*. Note that $\tilde{G}_X(y_1^N)$ is equivalent to the standard gaussian model scoring when *ACC* is an arithmetic mean.

The use of different time intervals enables us to discard or de-emphasize segments corresponding to abnormal events or segments poorly representative of the target speaker. The accumulation function proposed to take advantage of this segmentation is described in the next section.

3. FRAME PRUNING

A pruning procedure has been proposed in [9] where the 'neutral' frames (from which no particular speaker model emerges) are eliminated with a divergence measure. As far as our system is concerned, the underlying hypothesis is not the same. Pruning procedure is based on the assumption that the maximum likelihood scores resulting in correct identification are in general higher than the maximum likelihood scores resulting in incorrect identifications. In other words, when a segment is error-prone (i.e. when the true speaker is not identified on this segment), it is not due to a non-target speaker model matching well the speech segment, but rather to the true speaker model performing badly. We will see in section 4.3 that this assumption is warranted for the segment scores used in our experiments. Then, it is rather unlikely for non-target speaker models to achieve high log-likelihood scores.

3.1 Normalization

The likelihood scores of a segment must be normalized before pruning in order to make comparisons between segment scores meaningful. Actually, if the log-likelihood $\overline{G_X}(y_{t+1}^{t+T})$ obtained with a speech segment [t+1,t+T] is lower than the log-likelihood $\overline{G_X}(y_{k+1}^{k+T})$ obtained with a speech segment [k+1,k+T], it does not necessarily mean that segment [t;t+T] is less specific (of target speaker X) than segment [k;k+T]. Both segments convey different data and there is no basis for a meaningful comparison of their log-likelihoods.

Consequently, we propose to use a likelihood ratio as a normalized score. The denominator of the likelihood ratio is usually calculated using a collection of background speaker models. Different background speaker sets have been proposed in [8]: all other speakers, *top M* speakers, cohort speakers. Since our goal is not to study in detail the different normalization techniques, we will use the following normalized score [6] (equivalent to the *top 1* background speaker set of [8]):

$$l(y_{t+1},\ldots,y_{t+T}/X) = \frac{G(y_{t+1},\ldots,y_{t+T}/X)}{\max_{Z\neq X} G(y_{t+1},\ldots,y_{t+T}/Z)} \quad (5)$$

The numerator is the likelihood over the model belonging to speaker X, and the denominator is the maximum over all models not belonging to speaker X.

For convenience, we deal with the *minus-log likelihood ratio* rather than with the likelihood ratio itself. In that case, the normalized score becomes:

$$h(y_{t+1},\ldots,y_{t+T}/X) = -\log\ l(y_{t+1},\ldots,y_{t+T}/X) \quad (6)$$

which is equivalent to

$$h_X(y_{t+1}^{t+T}) = \max_{Z\neq X}\ \overline{G_Z}(y_{t+1}^{t+T}) - \overline{G_X}(y_{t+1}^{t+T}) \quad (7)$$

h is also called discriminant function [5] (p.52) since if h<0, speaker X scores higher than everyone else in the given segment and so speaker X is recognized on this segment; if h>0, the speaker recognized on the segment is not speaker X.

In our experiments, the normalizing speaker, for a given person, is chosen among the other speakers of the reference database, rather than among a completely separate group. Speakers are thus normalized by each other.

3.2 Pruning

Pruning can be achieved with the normalized scores; the modified minus-log-likelihood ratio of the whole test utterance of N frames is then:

$$\tilde{h}_X(y_1^N) = \arg\min_p\left[\frac{1}{p}\sum_1^p h_X(y_{tT+1}^{tT+T})\right]_{t\in[0\ldots n-1]} \quad (8)$$

In this case, we use the *p* lowest segment scores for each speaker, with *p<n* (*n* number of segments in the test utterance). We select the lowest scores because these are comparable to distances (minus log-likelihood); if the scores were log-likelihoods, we would keep the highest ones. We also note that the segments selected in the sum can vary from one speaker to the others.

4. EXPERIMENTS

4.1 Database and signal analysis

For our experiments, we have used TIMIT and NTIMIT databases. TIMIT [4] contains 630 speakers (438 male and 192 female), each of them having uttered 10 sentences. The NTIMIT database [7] was obtained by playing TIMIT speech signal through an artificial mouth installed in front of the microphone of a fixed handset frame and transmitting this input signal through a different telephone line for each sentence (local or long distance network).

The speech analysis module extracts filterbank coefficients in the following way: a Winograd Fourier Transform is computed on Hamming windowed signal frames of 31.5 ms (i.e. 504 samples) at a frame rate of 10 ms (160 samples). For each frame, spectral vectors of 24 Mel-Scale Triangular-Filter Bank coefficients (24 channels) are then calculated from the Fourier Transform power spectrum, and expressed in logarithmic scale. Covariance matrices and mean vectors are finally computed from these spectral vectors. These analysis conditions are identical to those used in [1] [2] and [3].

For TIMIT database, all 24 coefficients of the spectral vectors are kept. For NTIMIT, we remove the first 2 coefficients and the last 7 coefficients which are outside the telephone band (approximately 300-3400 Hz).

4.2 Training and test protocols

A common training/test protocol is used for all the experiments. In this protocol, training or test durations are rigorously the same for each speaker. Short durations are used (6s training and 3s test) in order to show the efficiency of the pruning procedure even when little speech material is available.

For the training of a given speaker, all 5 'sx' sentences are concatenated together and the first M samples corresponding to the training duration required (6s here) are selected. Consequently, a single reference pattern is computed from exactly the same number of samples for each speaker.

For the test of a given speaker, all 'sa' and 'si' sentences (5 in total) are randomly concatenated together and blocks of N samples corresponding to the test duration required (3s here) are extracted until there is not enough speech data available. Consequently, the test patterns are computed from exactly the same number of samples for each speaker.

All the tests are made within the framework of text-independent closed-set speaker identification using a 1-nearest neighbour decision rule.

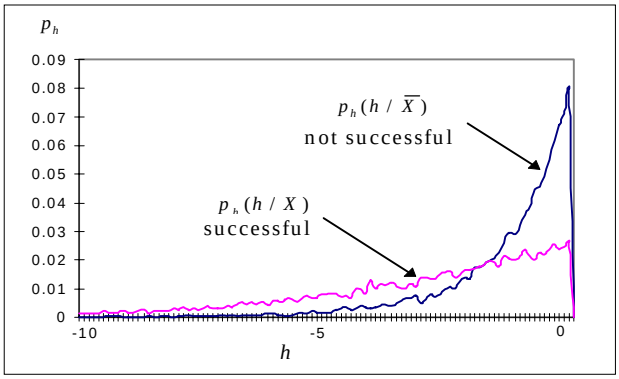

Figure 1. Density functions of $h_X(y_t)$ on the interval [-10,0]

4.3 Potential

The potential of the pruning procedure is illustrated by *Fig. 1* where the distributions of the normalized frame scores $h_X(y_t)$ are represented for speakers which score higher than everyone else in a given frame (i.e. negative values of $h_X(y_t)$). We distinguish 2 types of frames: frames on which the target speaker is recognized (successful frames) and frames on which a non-target speaker is recognized (not successful frames). The distributions of both classes are equivalent to the density functions of $h_X(y_t)$ and can be noted respectively $p_h(h/X)$ and $p_h(h/\overline{X})$.

We see that the frames may have lower minus log-likelihood ratios *h* for the true speaker ($p_h(h/X)$) than for non-target speakers ($p_h(h/\overline{X})$) which tends to prove the need of a pruning process to select the lowest values of *h* and thus eliminate error-prone frame scores.

4.4 Influence of the segment size *T*

We have investigated the influence of the number of frames *T* in a segment when no pruning process is performed (*p* = *n* total number of segments in a test utterance). Therefore, only the effect of normalization is studied. The speaker identification results obtained on a 63-speaker subset of TIMIT and NTIMIT (20 women, 43 men) are presented in *Tab. 1*.

T	300	150	100	50	30	20	10	5	1
p	1	2	3	6	10	15	30	60	300
TIMIT (%id.)	97.55	98.95	98.95	99.30	99.30	99.30	98.60	98.25	98.25
NTIMIT (%id.)	40.55	39.36	38.11	42.30	43.35	42.30	42.65	41.25	40.90

Table 1. Influence of the segment size 'T' (6s training/3s test - no pruning - 63 speakers - 286 tests)

We observe an optimum of the results for *T*=30 frames, i.e. when a normalization is made for each segment of 0.3s. In this case, the normalization alone leads to a significative improvement of the results on TIMIT compared to the standard

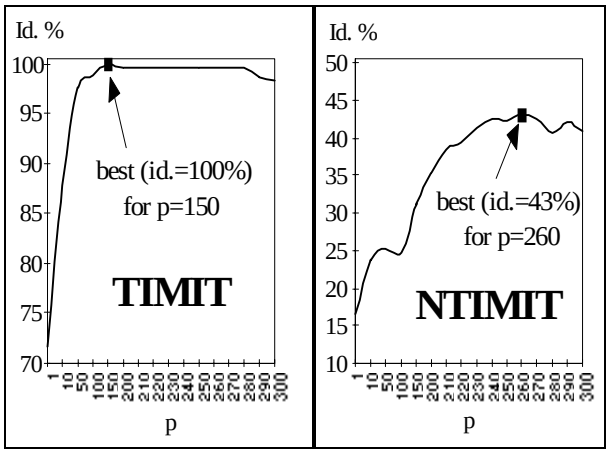

Figure 2. Influence of the number of frames kept 'p' (6s training/3s test - (300-p) frames pruned - T=1 - 63 speakers - 286 tests)

gaussian measure without normalization (T=300). A performance enhancement is observed on NTIMIT but it is not significative. These results confirm the usefulness of the normalization procedure for closed set speaker identification, already observed in [8].

4.5 Influence of the number of frames kept *p*

We have investigated the influence of the number of segments selected *p* when a segment is composed of a single frame (*T*=1). The results obtained with the same experimental conditions used in section 4.4 are reported in *Fig. 2*.

For both databases, optimum results are obtained when some frames are pruned. It shows that the selection of the information is important since some frames in a test utterance can contaminate the final score. Moreover, it is interesting to notice that a reasonably good performance is obtained on TIMIT when a single frame per speaker is kept (71.63% id.) that is to say when an extremely small amount of speech is used for each speaker to take the final decision !

5. VALIDATION

The optimal values of *p* and T obtained for 63 speakers on TIMIT and NTIMIT have been used to validate the benefit of the pruning procedure for speaker recognition. Speaker identification tests have been conducted on the 567 remaining speakers of TIMIT and NTIMIT. So the final test set is completely distinct from the tuning set from which the optimal values of *p* and T are evaluated. The identification results obtained are presented in *Tab. 2*. For both databases, the improvement of performances is significative which shows the interest of frame pruning for speaker recognition.

	BASELINE no norm. no-pruning	NORMALIZATION optimal T no-pruning	FRAME PRUNING norm. after each frame optimal *p*	IMPROVEMENT due to pruning / baseline
TIMIT	p=1;T=300	p=10;T=30	p=150;T=1	27% enhancement
Id. %	91.66	93.82	94.20	
NTIMIT	p=1;T=300	p=10;T=30	p=260;T=1	17% enhancement
Id. %	15.91	16.86	18.64	

Table 2. Validation of the pruning procedure on TIMIT and NTIMIT (6s training/3s test - 567 speakers - 2639 tests)

6. CONCLUSION

We have presented a frame pruning procedure for speaker recognition. It is shown, by results obtained, that this technique can significantly increase the performances of a speaker identification system. We intend to apply this method to refine the training of the speaker models. Another interesting issue would be to know systematically the phonetic label of rejected frames to say which specific part of a speech signal best identifies a speaker.

7. REFERENCES

[1] BESACIER, L., BONASTRE, J.F., Independent processing and recombination of partial frequency bands for automatic speaker recognition. *In Proc. International Conference on Speech Processing*, 26-28 August 1997. Seoul (Korea).

[2] BESACIER, L., BONASTRE, J.F., Subband approach for automatic speaker recognition: optimal division of the frequency domain. *In Audio- and Video-based Biometric Person Authentication*, Bigün, et. al. Eds., Springer LNCS 1206, 1997.

[3] BIMBOT, F., MAGRIN-CHAGNOLLEAU, I., MATHAN, L., Second-order statistical methods for text-independent speaker identification. *Speech Communication*, n°.17(1-2), August 1995.

[4] FISHER, W., ZUE, V., BERNSTEIN, J., PALLET, D., An acoustic-phonetic database. *JASA*, suppl. A, Vol. 81(S92). 1986.

[5] FUKUNAGA, K., *Statistical Pattern Recognition.* Second Edition, Academic Press, Inc., San Diego. 1990.

[6] GISH, H., SCHMIDT, M., Text independent speaker identification. *IEEE Signal Processing Magazine*, pp 18-32, October 1994.

[7] JANKOWSKI, C., KALYANSWAMY, A., BASSON, S., SPITZ, J. NTIMIT: A Phonetically Balanced Continuous Speech, Telephone Bandwidth Speech Database. *In Proc. ICASSP 90*, April 1990.

[8] MARKOV, K., NAKAGAWA, S., Frame level likelihood normalization for text-independent speaker identification using GMMs. *In Proc. ICSLP*, pp 1764-1767, 1996.

[9] VERGIN, R., O'SHAUGHNESSY, D., A Double Gaussian Mixture Modeling Approach to Speaker Recognition. *In Proc. Eurospeech 97*, Rhodes (Greece).