Frame Selection for Text-independent Speaker Recognition

Abedenebi Rouigueb, Malek Nadil and Abderrahmane Tikourt
Applied Mathematics Laboratory, Ecole Militaire Polytechnique, Algiers, Algeria

Keywords: Frame Selection, Frame Pruning, Speaker Identification, Speaker Recognition, Text-independent.

Abstract: In this paper, we propose a set of criteria for the selection of the most relevant frames in order to improve text-independent speaker automatic recognition (TISAR) task. The selection is carried out on the short term Cepstral feature vectors such as PLP and MFCC and performed at the front end processing level. The proposed criteria mainly attempt to select vectors lying far from the universal background model (UBM). Experiments are conducted on the MOBIO database and show that the selection allows an improvement in complexity (time and space) and in speaker identification rate, which is appropriate for real-time TISAR systems.

1 INTRODUCTION

Text-independent automatic speaker recognition (TIASR) task consists in verifying or in identifying the speaker identity using a segment of his speech where the utterance content is free (Beigi, 2011; Kinnunen and Li, 2010). Although a wealth of research works has addressed this problem throughout the last 40 years, it is still a challenging problem with many potential applications.

Cepstral short-term features extracted from short time frames of about 20-30 milliseconds in duration, such as Mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) coefficients (Kinnunen and Li, 2010), are widely used in TIASR systems. This choice is in part justified by the non-stationary nature of voice and the free speech assumption. The order between the feature vectors (e.g MFCC) is usually not utilized; the TIASR is merely treated as a recognition problem of a set, and not a sequence, of the acoustic vectors coming from the test utterance.

In typical TIASR systems, voice activity detection (VAD) (Benyassine et al., 1997) is first applied for removing non-speech frames. Unfortunately, due to the high variability and the non deterministic nature of noise, VAD cannot be perfectly achieved in all cases. VAD is more critical for non-stationary noise environments. Hence, some signal segments corresponding to noise and not to speech might be left. For further details about the challenges facing VAD techniques, refer to (Ramrez et al., 2004). Moreover, the noise cannot be completely filtered from the speech signal, either because the signal is too much altered or because the noise type is unknown. For useful discussions, see (Loizou, 2013).

In the TIASR context, different phonemes can be pronounced within a given utterance where each one produces about 5-10 Cepstral vectors. It has been shown that nasal consonants and vowels are more speaker-discriminative than the other phonemes like stops. Indeed, the speaker discrimination quality depends closely on the phonetic content. A quantitative assessment of phoneme groups for speaker recognition is given in (Eatock and Mason, 1994).

To recap, acoustic vectors resulting from i) non-speech segments, due to VAD failure; ii) severely corrupted speech segments; or iii) less-discriminative phonemes, are less relevant than those extracted from enhanced speech frames of discriminative phoneme group.

In light of the above-mentioned considerations, the TIASR task can be perceived as a classification problem of the acoustic vectors set resulting from the test utterance where they are involved with different degrees in the final decision.

In this paper, we aim at proposing selection methods of the most speaker discriminative acoustic vectors (frames). It is obvious that front-end frame selection leads to a significant reduction of the spatio-temporal complexity since the recognition process is achieved based only on a small subset of vectors. Our concern is focused here on the assessment of the influence of the amount of the selected data on the speaker

Rouigueb, A., Nadil, M. and Tikourt, A.
Frame Selection for Text-independent Speaker Recognition.
DOI: 10.5220/0006392100510057
In *Proceedings of the 14th International Joint Conference on e-Business and Telecommunications (ICETE 2017) - Volume 5: SIGMAP*, pages 51-57
ISBN: 978-989-758-260-8

recognition accuracy using several criteria. The performance is measured in terms of the selection task time and the speaker identification rate in relation to the percentage of the selected data.

The remainder of this paper is organized as follows. Section 2 presents related works and Section 3 discusses the utility of frame selection in TISAR systems. Section 4 shows the importance of the universal background model in the state-of-the-art of TIASR models. In Section 5 we present the proposed criteria, experimentations and results are discussed in Section 6. We highlight the improvements done in relation to our prior works in Section 7. Finally, conclusions and future directions are given in Section 8.

2 RELATED WORKS

In TISAR systems, a test utterance signal is split into a sequence of short-time frames and a features vector is extracted from each frame. Data reduction is often performed; it may be achieved by two complementary approaches: feature selection or vector (frame) selection. A large number of works have addressed the frame selection problem (also referred to as frame pruning). For instance, silence frame removal and VAD are usually applied at an early stage to remove a part of irrelevant frames. In addition to the spatio-temporal complexity reduction, an interesting frame pruning technique must not compromise the recognition performance. Beyond to silence removal and VAD pruning, we present in this section some interesting works found in the literature that have treated the frame pruning problem using different criteria.

A test utterance is divided into multiple frames (Besacier and Bonastre, 1998a) or multiple time-frequency blocks (Besacier and Bonastre, 1998b) and the final identification score is computed with a limited number of the obtained frames (blocks). A discriminant function, estimated for each speaker, is used to remove frames (or blocks) having a low log likelihood ratio score of the speaker model against the speakers' background model. Authors have reported that using only a 30% frame pruning can increase significantly the identification rate where the experiments are conducted on the TIMIT and NTIMIT corpora.

In (Kinnunen et al., 2006), the number of vectors of the utterance test is reduced by silence removable and pre-quantization (PQ) in order to speed up the identification process. Pre-quantization aims to keep a subset of vectors using different PQ techniques as random sub-sampling, averaging and decimation, see (Kinnunen et al., 2006) for more details. McLaughlin et *al*. have shown that the application of three simple PQ methods, prior to GMM (Gaussian Mixture Model) matching, allows to compressing the test sequence by a factor of 20:1 without compromising the verification accuracy (McLaughlin et al., 1999).

Recently, Almaadeed et *al*. have proposed a real-time text-independent speaker identification system where the consonant frames are filtered out and the identification is based on the formants extracted from vowels (Almaadeed et al., 2016).

3 FRAME SELECTION FOR TISAR SYSTEMS

If we make a deep analysis of the most popular accurate TIASR paradigms: ranging from the GMM-UBM baseline (Reynolds et al., 2000) to the state-of-the-art of the speaker verification I-vector concept (Dehak et al., 2011) via the GMM supervector (Campbell et al., 2006) and the joint factor analysis, one can conclude that:

1. All these models make use of short term acoustic vector, particularly MFCC, where the utterance vectors are seen as a set;
2. Some regions in the feature MFCC space are penalized more than others. For instance, in the GMM-UBM baseline system, MFCC vectors with a high density within the UBM class are disadvantaged in the likelihood decision ratio detector (Reynolds et al., 2000). In the I-vector paradigm, a GMM supervector, M, is first extracted using the Bayesian MAP adaptation (Campbell et al., 2006). Then, an I-vector of low dimension, w, is computed such that $M = m + Tw$ where T is a $N \times P$ rectangular mapping matrix and m is the GMM supervectors mean. There is an alignment between the entries of M and the mixture components centers of the GMM-UBM model. The significant dimension reduction ($N << P$) is feasible because several entries of the vector $M - m$ are either zeros, or correlated.

   So MFCC vectors with a high likelihood within the GMM-UBM components corresponding to the less-discriminative entries of M are penalized in the decision function.

The frame selection can be interesting for TIASR systems when the utterances duration is not very short. In such situation, we need efficient algorithms requiring low complexity in time and in memory space. For example, TIASR running on smartphones may be a typical case of use.

From a geometric point of view, the feature MFCC regions don't hold the same quantity of the speaker-dependent information. ALL accurate TIASR systems based on MFCC attempt, either in an explicit or an implicit way, to determine the importance of MFCC vectors according to their location. Therefore, for an absolute research purpose, proposing new selection criteria can be useful in order to develop new features and models by the segmentation of the MFCC space into regions according to their relevance.

4 UNIVERSAL BACKGROUND MODEL UTILITY

The concept of the universal background model is successfully used in (Reynolds et al., 2000), in which the verification task is seen as a test between the hypothesis H0: X is uttered by the claimed speaker against the alternative hypothesis H1: X is uttered by an impostor. A GMM, commonly referred as GMM-UBM model, is trained from a collection of data from a large number of expected speakers and it is used to fit the UBM density distribution. Few years later, the GMM-UBM model has been very successful when used in representing the speaker-independent information rather than the alternative hypothesis H1. The key idea consists in mapping a given utterance to a fixed-size GMM supervector.

The GMM-supervector of an utterance is derived via the MAP adaptation of the GMM-UBM distribution (Campbell et al., 2006). It has been shown that the best overall performance is from adapting only the mixture components means (centers) (Reynolds et al., 2000) compared to weights and covariance matrices. First, a GMM λ fitting the distribution of the utterance MFCC vectors is estimated via the MAP adaptation. Then, a GMM-supervector is obtained by the concatenation of the mixture components means of λ. Indeed, a GMM-supervector defines the overall location of the components of λ in relation to the GMM-UBM reference model.

A GMM-supervector, M, is decomposed into two components as follows $M = m + Tw$ in the i-vector model case. m itself is a GMM-supervector corresponding to the UBM class; it is discarded in the recognition process since it represents speaker-independent information.

GMM-UBM is used to make a mapping of an utterance X to a supervector such as GMM supervector, JFA factors, or i-vector. The location of the MFCC vectors of an utterance in relation to the UBM is so important in these new models. It is worth recalling that the angle between two supervectors is a relevant feature; many scoring functions are based on the cosine kernel (Dehak et al., 2011)

In this paper, we put forward the proposals that high density regions in the UBM class contain often MFCC vectors coming from several speakers. Hence, these regions are less discriminative since they tend to model common information as: non-speech segments, noisy speech, and non-discriminative phonemes. In this sense, we suggest to propose selection criteria by adopting the second interpretation which consists in using UBM to estimate the importance of MFCC regions.

5 PROPOSED CRITERIA

Let $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ be a set of N acoustic vectors extracted from a given utterance U, where each $\mathbf{x}_i$ is a D-dimensional cepstral vector (e.g MFCC, PLP). For the sake of simplicity, we propose to select a percentage of the vectors of X that maximize the proposed criteria because they have not the same duration in general. The proposed criteria are as follows.

1. Standard deviation (C_1)
   For a feature vector $\mathbf{x} = (x^1, ..., x^D)^T$, C_1 measures the dispersion of the different entries of $\mathbf{x}$, it is calculated as follows:

   $$C_1(\mathbf{x}) = \sqrt{\frac{1}{D}\sum_{i=1}^{D}(x^i - \overline{x})^2}, \qquad (1)$$

   where $\overline{x}$ is the mean value of $\mathbf{x}$ entries ($\overline{x} = \Sigma_{i=1}^{D} x^i / D$).
2. Euclidean distance (C_2)
   C_2 measures the Euclidean distance between the vector of interest $\mathbf{x}$ and the center of UBM, μ.

   $$C_2(\mathbf{x}) = \|\mathbf{x}, \mu\| = \sqrt{\Sigma_{i=1}^{D}(x^i - \mu^i)^2} \qquad (2)$$
3. Probabilistic distance (C_3)

   A GMM, λ_{ubm}, with K mixture components is trained to approximate the UBM distribution. $C_3(\mathbf{x})$ is inversely proportional to the likelihood of UBM. We suggest:

   $$C_3(\mathbf{x}) = -P_{\lambda_{ubm}}(\mathbf{x}) \qquad (3)$$

   We have experimented the variants $C_3^{(1)}$, $C_3^{(4)}$, $C_3^{(32)}$, $C_3^{(64)}$, $C_3^{(128)}$, $C_3^{(256)}$ by setting K to 1, 4, 32, 64, 128, and 256, respectively, in λ_{ubm}.

4. Probabilistic distance with component selection (C_4)

   The time cost of the evaluation of C_3 linearly increases with the number of components in λ_{ubm}, K, because we need to compute the density of $\mathbf{x}$ in each one of them.

   Each component of λ_{ubm} corresponds to a specific sound or phoneme group. Obviously, some components better describe speaker-dependent information than others. C_4 constitutes an improvement of C_3, where the key idea consists in composing a new GMM, λ'_{ubm}, only from the less-discriminative components of λ_{ubm}. Hence, we gain in selection speed and possibly in accuracy.

   Let $\lambda_{ubm} = \{w_i, m_i, \Sigma_i\}_{i=1..K}$ be a K-component GMM-UBM which is estimated from the speech of S speakers. First, we compute a confusion matrix H where each cell $H(s,k)$ represents the portion of the likelihood of s which is expressed by the component k. Formally, we have:

   $$H(s,k) = \frac{w_k\, P(N(m_k,\Sigma_k);Xs)}{\sum_{i=1}^{K} w_i\, P(N(m_i,\Sigma_i);Xs)}, \qquad (4)$$

   Xs is the data of speaker s. The sum of each row of H is equal to 1. The standard deviation of the column of index k measures the variability of the component k. Low standard deviation means that the speakers' likelihood values are close to each other and that the component k is consequently less-discriminative.

   In experiments, we have built λ'_{ubm} from a quarter of the less-discriminative components of λ_{ubm}. We have experimented the variants $C_4^{(4)}$, $C_4^{(32)}$, $C_4^{(64)}$, $C_4^{(128)}$ by setting K to 4, 32, 64, 128, respectively. For example, in $C_4^{(32)}$, λ'_{ubm} contains only the 08 less-discriminative components of the 32 components forming λ_{ubm}. Thus, we suggest:

   $$C_4(\mathbf{x}) = -P_{\lambda'_{ubm}}(\mathbf{x}) \qquad (5)$$

6 EXPERIMENTS

6.1 Dataset

All of the results we report are on the MOBIO database (McCool et al., 2012). Following the same spirit as the NIST SRE, the Biometric Group at the Idiap Research Institute organized the evaluation on text-independent speaker recognition.

MOBIO is a bimodal (audio and video) database recorded from 152 persons, 100 males and 52 females with both native and non-native English speakers. For each individual, 12 sessions were captured where 192 utterances are recorded by mobile phone (NOKIA N93i). MOBIO is a challenging database since the data is acquired on mobile devices possibly with real noise, and the speech segments can be very short (less than 02 sec). The average speech duration of MOBIO phrases is around 08 sec. MOBIO is designed so that it contains realistic and common environmental variations associated with the usage of mobile devices. More details on this dataset could be found in (Khoury et al., 2013).

6.2 Methodology

We propose to evaluate the performance of the proposed criteria inside a speaker identification system. It is clear that useful information for the speaker identification task, is also useful for the speaker verification task and vice-versa.

The vectors selection module is integrated in the identification system as illustrated in Figure 1. A percentage of cepstral vectors of each training or test utterance that maximize the selection criterion are used for identification. The selection module may use the parameters of the GMM-UBM model.

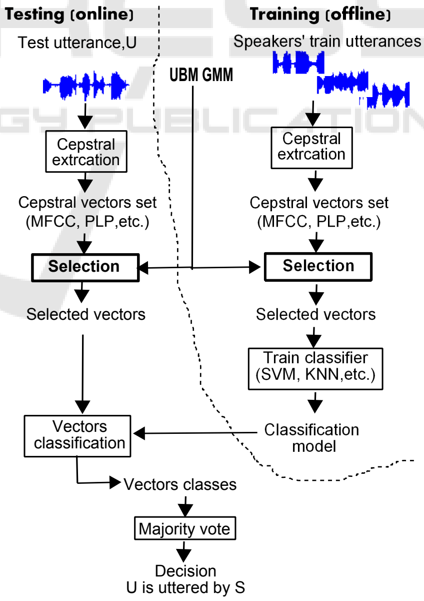

Figure 1: Cepstral data selection for speaker identification.

The objective of the experimentation is to study the identification performance as function of the following parameters.

1. The percentage of selected data, $\theta \in [1,..,100]$;
2. The selection criterion, $C \in \{C_1, C_2$, variants of C_3 and $C_4\}$

3. The classifier used, $Cl \in \{$ SVM , KNN $\}$.

For a given combination (θ, C, and Cl), an identification experiment is carried out using 20 speakers where 10 and 30 utterances are used in test and training (resp.) for each speaker. Speakers and utterances involved in experiments are drawn randomly from the whole MOBIO corpus. Then, the following performance measures are computed:

- The identification rate : $\tau = \frac{number\ of\ correct\ identification\ trials}{number\ of\ alltrials = 20*30 = 600}$;
- The time required for vectors selection of the test utterances, T_{sel} ;
- The time required for vectors classification of the test utterances, T_{cl}.

6.3 Experimental Results & Discussions

In this section, we present experiments and results obtained in order to observe the influence of the selection parameters over the identification performance. Each identification experiment is conducted as shown in subsection 6.2.

The first set of experiments examine the variation of the identification rate, τ, as function of the selection criterion, C, and percentage, θ. Used features consist of 19 MFCCs acoustic vector, computed with a frame shift of 10 ms and a frame size of 25 ms. The identification rate curves when SVM (resp. KNN) is applied are depicted in Figure 2 (resp. 3).

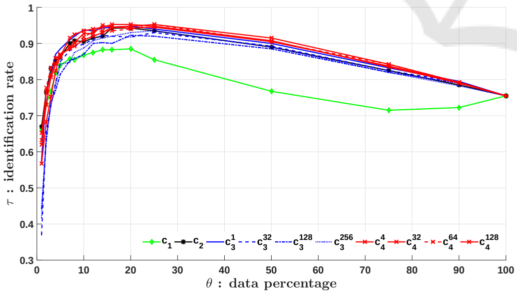

Figure 2: Accuracy as function of criteria and data percentage using SVM classifier.

While the difference between criteria in terms of τ for both figures is not large, C_4 variants performed slightly better than the rest. This seems logical since C_4 attempts to keep only vectors having low likelihood in the less-discriminative mixture components of the GMM-UBM model.

The high rates obtained by C_2 and C_3^1 are surprising and prove that the GMM-UBM distribution has roughly a hyper-spherical form which is much dense around the center.

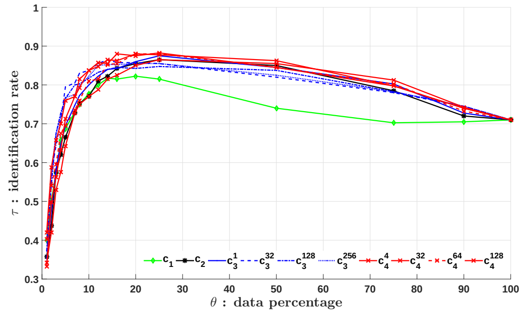

Figure 3: Accuracy as function of criteria and data percentage using KNN classifier.

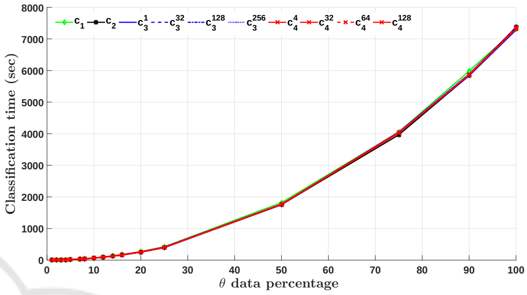

Figure 4: Classification time, T_{cl},as function of data percentage, θ.

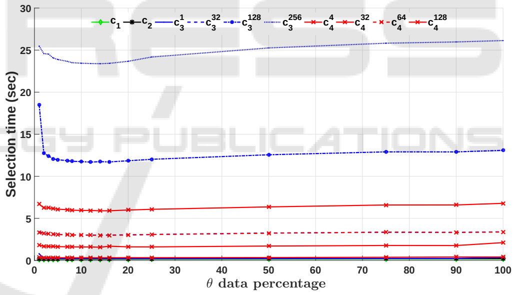

Figure 5: Selection time, T_{sel},as function of data percentage, θ.

When examining the variations of τ of all criteria in both figures 2 and 3, one can clearly see three distinct regions of θ, as follows:

a) $\theta \in [0\% - 5\%]$: the identification rates obtained are too poor because the selected data are not sufficient in amount for the recognition task;

b) $\theta \in [5\% - 25\%]$: the highest rates are achieved in this interval;

c) $\theta \in [25\% - 100\%]$: a continuous deterioration of τ is seen by the increasing of θ value. Acoustic vectors minimizing the proposed criteria don't hold enough of speaker-dependent information, their incorporation in the recognition task leads to a decreasing in performance.

In Figure 4, we show the classification time (for

600 identification test trials) using SVM. The time curves are quite identical for all criteria and seem to have a quadratic augmentation. Indeed, the classification time depends mainly on the data percentage, θ. In Figure 5, we show the selection time. As expected, the fastest criteria are C_1, C_2, C_3^1, C_4^1 because few operations are needed to evaluate them. The test time includes principally selection and classification. Therefore, for a global comparison, C_2,C_3^1,C_4^1, represent a good trade-off between accuracy and complexity. We deduce that taking θ in [10% 25%] is interesting in the MOBIO corpus case.

Table 1: Best identification rates using PLP, MFCC+Δ features.

	PLP		MFCC+Δ	
	SVM	KNN	SVM	KNN
τ	0.9550	0.8900	0.9775	0.8550
C^*,θ^*	C_3^1,3%	C_3^1,3%	C_4^{256},18%	C_3^{256},15%

Table 2: The influence of the intersession variability.($C = C_3^1$, classifier=SVM).

sessions	τ^*,θ^*	sessions	τ^*,θ^*
1	0.990, 07%	7	0.990, 10%
2	0.990, 20%	8	0.990, 07%
3	0.995, 04%	9	1.000, 04%
4	0.990, 12%	10	0.990, 10%
5	1.000, 07%	11	1.000, 16%
6	0.990, 14%	12	1.000, 07%
all(1-12)	0.95, 16%		

The same set of experiments is achieved on the PLP, MFCC+delta features. We have noticed the same behavior for both cases where the best results are obtained by setting $\theta \leq 20\%$, see Table 1.

The last set of experiments aims to explore the influence of selection towards the session differences. The intersession variability is well-known to be a hard problem; several works focused on this challenge could be found in the literature.

We recall that 12 different sessions are recorded for each person in MOBIO. We can notice in Table 2 that τ, when it is computed for each session separately (training and test utterances coming from the same session), is almost equal to 1, this corresponds to the first 12 cases. On the other hand, τ of the 'all sessions' case (majority of utterances come from different sessions) is significantly low, see the last line in Table 2. This observation proves that after achieving data selection, the majority of the identification failures are due to the multi-session problem. Therefore, session compensation is still needed to address this problem.

7 PRIOR WORK

In this paper, we attempted to extend our previous work (Tikourt et al., 2015) dealing with MFCC selection. In short, the major improvements are summarized in Table 3.

Table 3: Improvements.

	previous work	current work
features	MFCC	MFCC, PLP, ΔMFCC
sel. criteria	$C_1,..,C_3$	$C_1,..,C_4$
classifier	SVM	SVM, KNN
# speaker	10	20
# training utter.	10	10
# test utter.	10	30
session exploration	no	yes

In addition to SVM, we propose to apply the KNN classifier, because this last one is non-parametric, robust and able to separate classes with non-linear complex boundaries. The suggested improvements aim partly to consolidate results about MFCC selection obtained in our previous work (Tikourt et al., 2015).

8 CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have described a set of criteria proposed to select the most relevant short-term feature vectors for the text-independent recognition task.

A universal background model is used for representing the speaker-independent information, and hence it can be used as a framework for the selection purpose. The general idea consists in selecting vectors having a low likelihood in the UBM class.

Speaker identification experimental tests on the MOBIO corpus are presented. Results show that the relevant speaker information is contained in less of 20% of data maximizing the criteria. Not only using the vectors that minimize the criteria increases the complexity in time and space, but also reduces the identification rate.

The findings of this study show that the distance Euclidean from the UBM center, and the minus-likelihood in the one-component Gaussian of UBM are efficient. This supports the idea that UBM has approximately a hyper-spherical distribution form, such as a multivariate normal distribution with equal variances.

It is obvious that an efficient frame pruning speeds up the recognition computational process since a small percentage of data is kept for speakers model matching. Nevertheless, the recognition performance is enhanced only if a good trade-off between the frame pruning and the speaker modeling is made. On the one hand, pruning the irrelevant frames causes a loss in speaker information, but it makes easier the task of fitting the speakers' models to data. On the other hand, using all the frames preserves the entire speaker information, but it makes the model estimation inaccurate and more complex. This work and the review of the literature have led us to conclude that for TISAR systems an efficient frame pruning, if it is combined with a suitable modeling, may speed up significantly the recognition task without too much compromising (even improving) the accuracy. In this optic, frame pruning is an important approach to design real-time TISAR systems.

The majority of related works attempt to remove some kinds of the irrelevant frames using specific criteria based on the silence, the noise, the phonetic information, or the correlation between successive frames. The main contribution of this work consists in applying the UBM model to prune all the irrelevant frames at once whatever the kind.

To further our research we plan to use this finding inside a verification TISAR system. Moreover, developing efficient frame pruning techniques could be used as a basis to propose new features (e.g supervectors) or models by setting more of importance to vectors maximizing the selection criteria.

REFERENCES

Almaadeed, N., Aggoun, A., and Amira, A. (2016). Text-independent speaker identification using vowel formants. *Journal of Signal Processing Systems*, 82(3):345–356.

Beigi, H. (2011). *Fundamentals of Speaker Recognition*. Springer Publishing Company, Incorporated.

Benyassine, A., Shlomot, E., Su, H. Y., Massaloux, D., Lamblin, C., and Petit, J. P. (1997). Itu-t recommendation g.729 annex b: A silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications. *Comm. Mag.*, 35(9):64–73.

Besacier, L. and Bonastre, J. F. (1998a). Frame pruning for speaker recognition. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 765–768 vol.2.

Besacier, L. and Bonastre, J. F. (1998b). Time and frequency pruning for speaker identification. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)*, volume 2, pages 1619–1621 vol.2.

Campbell, W. M., Sturim, D. E., and Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Process. Lett.*, 13(5):308–311.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798.

Eatock, J. P. and Mason, J. S. D. (1994). A quantitative assessment of the relative speaker discriminating properties of phonemes. In *Proceedings of ICASSP '94: IEEE International Conference on Acoustics, Speech and Signal Processing, Adelaide, South Australia, Australia, April 19-22, 1994*, pages 133–136.

Khoury, E., Vesnicer, B., and Franco-Pedroso, e. a. (2013). The 2013 speaker recognition evaluation in mobile environment. Idiap-RR Idiap-RR-32-2013, Idiap.

Kinnunen, T., Karpov, E., and Franti, P. (2006). Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):277–288.

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

Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition.

McCool, C., Marcel, S., Hadid, A., and et *al.* (2012). Bi-modal person recognition on a mobile phone: using mobile phone data. Idiap-RR Idiap-RR-13-2012, Idiap.

McLaughlin, J., Reynolds, D. A., and Gleason, T. P. (1999). A study of computation speed-ups of the gmm-ubm speaker recognition system. In *EUROSPEECH*. ISCA.

Ramrez, J., Segura, J. C., Bentez, C., Torre, A. D. L., and Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42:3–4.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, page 2000.

Tikourt, A., Rouigueb, A., and Djeddou, M. (2015). Efficient data selection criteria for speaker recognition. In *3rd Inter. Conf. on Signal, Image, Vision and their Applications*, Guelma, Algeria.