On the Fusion of Dissimilarity-Based Classifiers for Speaker Identification

Tomi Kinnunen, Ville Hautamäki, Pasi Fränti

Department of Computer Science
University of Joensuu, Finland
{tkinnu,villeh,franti}@cs.joensuu.fi

Abstract

In this work, we describe a speaker identification system that uses multiple supplementary information sources for computing a combined match score for the unknown speaker. Each speaker profile in the database consists of multiple feature vector sets that can vary in their scale, dimensionality, and the number of vectors. The evidence from a given feature set is weighted by its reliability that is set in *a priori* fashion. The confidence of the identification result is also estimated. The system is evaluated with a corpus of 110 Finnish speakers. The evaluated feature sets include mel-cepstrum, LPC-cepstrum, dynamic cepstrum, long-term averaged spectrum of /A/ vowel, and F0.

1. Introduction

Speaker individuality is a complex phenomenon, where different supplementary information sources contain a part of evidence of the speaker identity. The individual speaker characteristics occur both at the lexical, segmental and prosodic levels [11]. At the lexical level [15] this is reflected, for instance, in usage of certain word patterns. At the segmental level, speaker differences occur at the acoustic differences of phoneme realizations that arise from physiology and anatomy of the voice production organs. Prosodic speaker characteristics are reflected in the usage of pitch, stress and timing.

Extraction of individual characteristics is realized by measuring *acoustic parameters* or *features* from the speech signal. Commonly used features in automatic speaker recognition systems include mel-cepstrum, LPC-cepstrum [1], line spectral frequencies [10], subband processing [6], dynamic cepstral parameters [14], and prosodic parameters [15].

Spectral parameters alone, especially the cepstrum with its variants, have shown good performance in speaker recognition. However, cepstrum carries only one source of evidence. To achieve better recognition accuracy, several supplementary information sources should be used.

The idea of using multiple features in speaker recognition is not new. A well-known data fusion strategy is to concatenate the cepstral vectors with their delta- and delta-delta cepstra into a long feature vector [1]. Also the fundamental frequency has been used in addition with the cepstral vectors to improve recognition accuracy. In general, vector concatenation is termed as *classifier input fusion* [12].

Although classifier input fusion is simple to implement and works reasonably well, it has a few shortcomings. Firstly, the feature space formed by concatenation of different features is somewhat superficial. The higher the dimensionality of the space becomes, the less and less effect a single feature has to the overall match score. Also, fusion becomes difficult if the feature is missing (e.g. F0 for unvoiced sounds) or it should be computed with a different frame rate.

Another way of performing data fusion is to combine different classifiers. In *classifier output fusion*, each individual data source is modeled separately, and the outputs of the individual classifier scores are combined to give the overall match score. For instance, output fusion of the cepstral and delta-cepstral features has been performed using VQ codebooks [14] and Gaussian mixture models [12] as the individual classifiers.

Slomka & al. [12] compared input and output fusion for the mel-cepstrum and corresponding delta features. They found out that the output fusion performed consistently better. Furthermore, they demonstrated that the computational complexity for the input fusion is higher than that of the output fusion.

Classifier output fusion is, with to many respects, a flexible combination strategy. For instance, it enables the same data source to be modeled by several different classifiers. In [9], a committee of five learning vector quantization (LVQ) networks with different network structures was applied. The combination was done with majority voting rule.

The main objective of this paper is to design the fusion strategy such that evidences from diverse data sets could be combined in a coherent way. Problems arise when the data sources differ in (1) the number of features (dimensionality), (2) the number of measurements, (3) the scales. Furthermore, a model that works well for one data source might not be good to model another feature. Thus, each individual feature stream should be modelled with the most suitable model for that stream. The proposed classifier is invariant to different scales of feature sets, their dimensionality, and the number of measurements. For each feature set, an *a priori* weight is set based on the reliability of the feature set.

This work was carried out in co-operation with the Department of Phonetics at the University of Helsinki as a part of larger speaker recognition project [5]. To be reliable in, for instance, realistic forensic uses, speaker recognition should be based on many parameters instead of only spectral parameters. Forensic speech samples often suffer from different types of noises and distortions, and therefore, supplementary identity cues should be used to give a joint decision. The combination of supplementary evidences from diverse feature sets, however, is not a straightforward task. In this paper, we report the structure of the fusion system we designed for the use of this project. The experiments show that using multiple feature sets together improves recognition accuracy.

2. The structure of the system

The structure of the proposed data fusion system is shown in Fig. 1. The *profile* of each of the registered N speakers $S(i),\ i = 1, \ldots, N$, consists of M distinct models, $S(i) = \{S_1(i), \ldots, S_M(i)\}$. Each of the models consists of a set of feature vectors. For each model, there is a correspond-

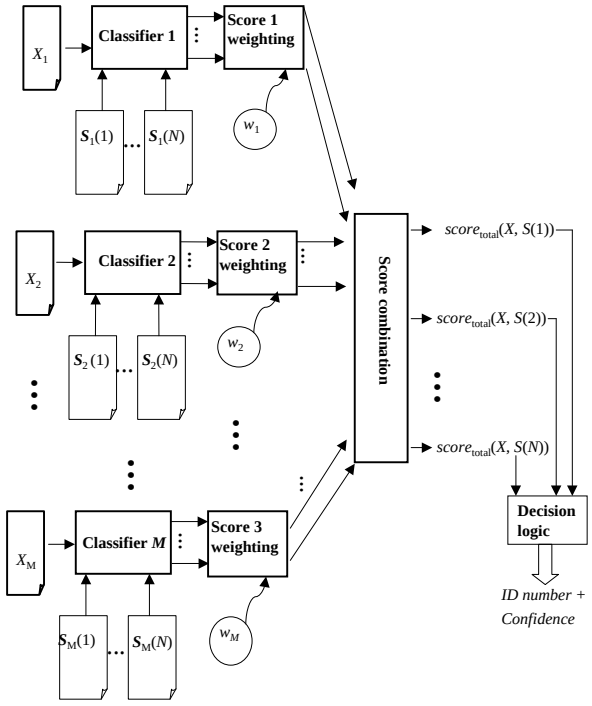

Figure 1: *Structure of the proposed system.*

ing *sub-classifier* or *expert*. Given an unknown speaker profile $X = \{X_1, \ldots, X_M\}$, each of the experts j computes a match score $score(j, i)$ for each speaker i. The match score $score(j, i)$ indicates the degree of similarity (or dissimilarity) between point sets X_j and $S_j(i)$.

The individual expert outcomes $score(j, i), j = 1, \ldots, M$ are weighted by *a priori* weights $w(j)$ that indicate the reliability of the expert. The weighted match scores from the different experts are then combined into a single match score $score_{\text{total}}(X, S(i))$ that indicates the degree of similarity (or dissimilarity) between the speakers X and $S(i)$. The decision is given by returning the ID number of the most similar speaker to X. The confidence of the decision is also estimated based on the spread of the distribution of the match scores from different speakers.

2.1. Sub-classifiers

For simplicity, we will use dissimilarity-based classifiers for all feature sets. For each speaker, the individual feature sets are modeled by codebooks [10, 6, 13] generated by clustering the feature vectors of that feature set by randomized local search algorithm [3].

Dissimilarity of point sets sets X_j and $S_j(i)$ is computed by the average quantization distortion:

$$D(j,i) = \frac{1}{|X_j|} \sum_{\vec{x} \in X_j} \min_{\vec{y} \in S_j(i)} \|\vec{x} - \vec{y}\|^2, \qquad (1)$$

where $|X_j|$ denotes the cardinality of X_j and $\|.\|$ denotes the Euclidean norm. The match score for the sub-classifier is computed as normalized distortion:

$$score(j,i) = \frac{D(j,i)}{\sum_{k=1}^{N} D(j,k)}. \qquad (2)$$

In other words, the distortion of each speaker within the sub-classifier is normalized by the sum of the distortions from all speakers within that sub-classifier. This ensures that $0 \leq score(j,i) \leq 1$. In this way, the outputs of the individual classifiers are in the same order of magnitude regardless of the dimensionality or the number of vectors.

2.2. Fusion strategy

There are several options for combining the outputs from the sub-classifiers [2, 7]. Kittler & al. [7] compared several commonly used fusion criteria in the context of of probabilistic classifiers. Their theoretical and experimental results indicated that the *sum rule* is most resilient to estimation errors. Therefore, we define the combination rule as the weighted sum:

$$score_{\text{total}}(X, S(i)) = \sum_{j=1}^{M} w(j)\, score(j,i), \qquad (3)$$

where $w(j)$ is the weight for the feature set j. The weights are normalized such that $\sum_{j=1}^{M} w(j) = 1$, which allows the weights to be interpreted as relative importances. For instance, if there are two feature sets and we set $w(1) = 0.2$ and $w(2) = 0.8$, then the second set gets four times more weight in the fusion compared to the first one.

2.3. Decision and confidence estimation

The identification decision is the speaker i^* which produces the smallest combined score:

$$i^* = \arg \min_{0 \leq i \leq N} score_{\text{total}}(X, S(i)). \qquad (4)$$

We also estimate the *confidence* of the decision. Intuitively, one should expect high confidence if the selected speaker is very distinctive, i.e. the scores for all other speakers are significantly higher. On the other hand, if there exists another speaker that is close to i^*, the decision is more uncertain. Based on this idea, we define the confidence as

$$c = 1 - \frac{score_{\min}}{score_{\min 2}}, \qquad (5)$$

where $score_{\min}$ and $score_{\min 2}$ are the scores for the nearest and second nearest speakers, respectively.

2.4. Determination of the weights

We consider two ways of determining the weights in Eq. (3). In the first approach, we apply a separability criterion for the within- and between-speaker distance scores within each feature set. The separability of the distributions is computed by the Fisher's criterion [4]:

$$F = \frac{(\mu_w - \mu_b)^2}{\sigma_w^2 + \sigma_b^2}, \qquad (6)$$

where μ_w, μ_b and σ_w^2, σ_b^2 are the means and variances of the two distributions, respectively. The Fisher's criterion gives a high value if the two distribution are well-separated.

In the second approach, we use exhaustive search to find the optimum weight combination. In other words, the performance of the system is evaluated for every weight combination, and the best weight combination is selected. For a small number of feature sets this approach can be applied.

Table 1: Summary of the data sets.

	Dimensionality	Vectors	Range
MFCC	16	499	[-102.9, 48.4]
△-MFCC	16	499	[-12.7,13.4]
△△-MFCC	16	499	[-5.5, 6.5]
LFCC	20	1990	[-18.1,48.4]
LTAS	513	1	[-25.6, 57.6]
F0	1	469	[57.9, 323.0]

Table 2: Performances of the subclassifiers.

	Error rate	Avg. confidence
MFCC	6.36 %	0.14
△-MFCC	52.72 %	0.05
△△-MFCC	46.36 %	0.04
LFCC	46.36 %	0.10
LTAS	5.45 %	0.53
F0	93.64 %	0.35

3. Experiments

3.1. Corpus description

The test material consists of 110 native Finnish speakers from various dialect regions in Finland [5]. The recordings were done in a silent environment by a professional reporter C-cassette recorder. The data was digitized using 44.1 kHz sampling frequency with 16 bits per sample. All speakers read the same material which was divided into training and evaluation sets of length 10 seconds both.

3.2. Acoustic measurements

The original acoustic measurements as provided by the University of Helsinki consisted of four data sets [5]: fundamental frequency (F0), long-term averaged spectrum (LTAS) for vowel /A/ , linear frequency cepstral coefficients (LFCC) and mel-cepstral coefficients (MFCC). We furthermore added the dynamic cepstrum parameters (△-MFCC, △△-MFCC) due to their popularity in automatic speaker recognition systems.

The data sets are summarized in Table 1. From this table, we can see that input fusion would be impossible due to the diversity of the data sets. The fusion system enables using arbitrary feature sets together.

3.3. Sub-classifier performance

First, the performances of each feature set alone were evaluated. After some experimentation, we fixed the model sizes as follows. For MFCC, LFCC, △-MFCC and △△-MFCC the models consist of 100 code vectors. For F0, the model consists of 5 code vectors. For LTAS, the model consists of, by definition, one long vector containing 513 averaged subband outputs from different instances of /A/ vowels.

The performances of the individual data sets are summarized in Table 2 for segment length 1.8 seconds. Both the identification error rate and average confidence for the correctly classified speakers are shown.

We found out that in general increasing the model size and the test segment length improves recognition results. An exception was F0, for which the behaviour was somewhat inconsistent with respect both to the model size and to the test segment length. From the six sets, MFCC and LTAS performed best and F0 worst.

Notice that the confidences do not go in parellel with the recognition rates. For instance, F0 gives poor identification result but the confidence for the correctly classifier speakers is higher than that of MFCC, for instance.

3.4. Fusion of data sources

Since the fundamental idea of the fusion is that the classifiers could complement each others results, the fusion of correlated classifiers is not reasonable. In other words, if two classifiers misclassify the same speakers, there is little gain in combining their outputs; in fact, the results may even get worse. To attack this potential problem, we computed the correlations between the classifier score outputs which are listed in Table 3.

We can see from Table 3 that LFCC is highly correlated with MFCC. This is an expected result, since both of them describe essentially the same quantity, spectral shape. Also, dynamic cepstral parameters are highly correlated with each other, which can be explained by the method they are computed: △△-MFCC is merely a differenced version of △-MFCC.

From the six data sets, LTAS and F0 are least correlated with the other feature sets. Based on these observations, we selected MFCC, LTAS and F0 for the evaluation of data fusion. The results for a test segment of length 1.8 seconds for the two best sub-classifiers and the fusion are compared in Table 4. It can be seen that by combining the data sets, the error rate is halved. This shows that the fusion strategy works as designed.

3.5. Weight selection

Next we study the effect of the weight selection. The results for equal weights, Fisher's criterion, and exhaustive search are compared in Fig. 2 for different input segment lengths.

Figure 2 indicates that the selection of weights has some importance. With exhaustive search, we can find the optimum weight combination for given model size and test segment length. However, this is computationally intensive approach and furthermore, the weights computed in this way do not give any insight into data sets themselves. Thus, the Fisher's criterion seems more appropriate choice for practical use. Both of these approaches outperform the equal weights case, which suggests that the feature sets, indeed, have unequal discrimination powers (reliability).

We continue by fixing the weights according to Fisher's criterion and examine what is the effect of excluding the best feature set, LTAS. The results are compared with MFCC in the Fig. 3. We observe that excluding LTAS increases error rate. Therefore, the gain in the fusion is mostly due to LTAS feature set. Fusion without LTAS is close to the results obtained us-

Table 3: *Correlations of the feature sets.*

	MFCC	LFCC	LTAS	Δ-MFCC	ΔΔ-MFCC
MFCC					
LFCC	0.88				
LTAS	0.19	0.13			
Δ-MFCC	0.72	0.62	0.14		
ΔΔ-MFCC	0.69	0.62	0.10	0.94	
F0	0.31	0.09	0.02	0.25	0.20

Table 4: Comparison of the two best subclassifiers and the data fusion.

	Error rate	Avg. confidence
LTAS	5.45 %	0.53
MFCC	6.36 %	0.14
Fusion	2.72 %	0.19

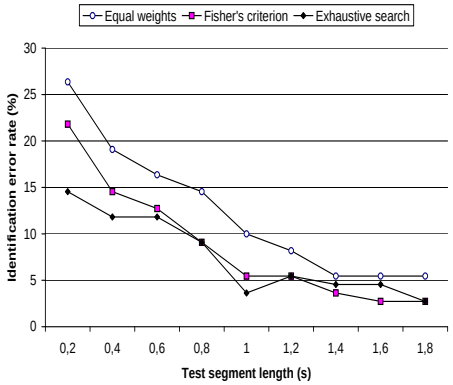

Figure 2: *Comparison of weight selection.*

ing MFCC alone. For very short segments, the data fusion still improves recognition accuracy.

4. Conclusions

Information fusion of diverse data sets is a difficult task. We have evaluated the performance of classifier output fusion for multiparametric speaker identification in the case of dissimilarity-based classifiers. The results indicate that by using multiple uncorrelated feature sets, the recognition performance of the fusion system is better than any of the sub-classifiers alone.

5. Acknowledgements

This work was partially carried out under the project "The Joint Project Finnish Speech Technology" supported by the National Technology Agency (agreements 40285/00, 40406/01, 40238/02) and titled "Speaker Recognition" (University of Helsinki ProjNr 460325).

6. References

[1] J. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, 85(9), pp. 1437-1462, 1997.

[2] R.P.W. Duin, "The Combining Classifier: To Train Or Not To Train?," *Proc. 16th International Conference on Pattern Recognition (ICPR 2002)*, Quebec City, Canada, pp. 765-770, 2002.

[3] P. Fränti, J. Kivijärvi, "Randomized Local Search Algorithm for the Clustering Problem," *Pattern Analysis and Applications*, 3(4), pp. 358-369, 2000.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.

[5] A. Iivonen, K. Harinen, M. Horppila, L. Keinänen, J. Kirjavainen, H. Liisanantti, E. Meister, L. Perälä, L. Tuuri, L. Vilhunen, "Development of a Multiparametric Speaker Profile for Speaker Recognition," manuscript, accepted for publication in *The 15th Int. Congress on Phonetic Sciences (ICPhS 2003)*, Barcelona, Spain, 2003.

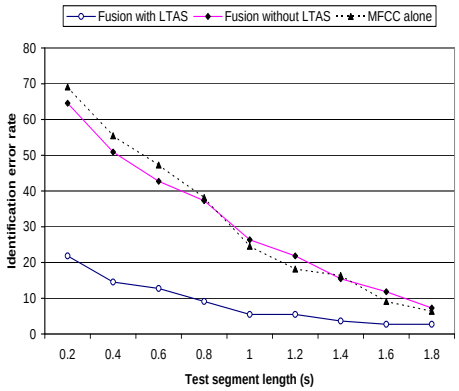

Figure 3: *Excluding the best feature set (LTAS).*

[6] T. Kinnunen, "Designing a Speaker-Discriminative Adaptive Filter Bank for Speaker Recognition," *Proc. ICSLP 2002*, pp. 2325-2328, Denver, USA, 2002.

[7] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3), pp. 226-239, 1998.

[8] K.P. Markov, S. Nakagawa, "Text-Independent Speaker Recognition Using Multiple Information Sources," *Proc. ICSLP 1998*, pp. 173-176, Sydney, Australia, 1998.

[9] V. Moonasar, G.K. Venayagamoorthy, "A Committee of Neural Networks for Automatic Speaker Recognition (ASR) Systems," *Proc. Int. Joint Conference on Neural Networks*, pp. 2936-2940, Washington DC, USA, 2001.

[10] P.C. Nguyen, M. Akagi, T.B. Ho, "Temporal Decomposition: A Promising Approach to VQ-Based Speaker Identification," manuscript, accepted for publication in *ICASSP 2003*, Hong Kong, 2003.

[11] P. Rose, *Forensic Speaker Identification*, Taylor & Francis, London, 2002.

[12] S. Slomka, S. Sridharan, V. Chandran, "A Comparison of Fusion Techniques in Mel-Cepstral Based Speaker Idenficication," *Proc. ICSLP 1998*, Sydney, Australia, 1998.

[13] F.K. Soong, A.E. Rosenberg, B.-H. Juang, and L.R. Rabiner, "A Vector Quantization Approach to Speaker Recognition," *AT & T Technical Journal*, 66, pp. 14-26, 1987.

[14] F.K. Soong and A.E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, 36(6), pp. 871-879, 1988.

[15] F. Weber, L. Manganaro, B. Peskin, E. Shriberg, "Using Prosodic and Lexical Information for Speaker Identification," *Proc. ICASSP 2002*, pp. 141-144, Orlando, USA, 2002.