

Selective Fusion for Speaker Verification in Surveillance

Yosef A. Solewicz^{1,2} and Moshe Koppel¹

¹ Dept. of Computer Science, Bar-Ilan University, Ramat-Gan, Israel

² Division of Identification and Forensic Science, Israel National Police, Jerusalem, Israel
solewicz@013.net.il
koppel@netvision.net.il

Abstract. This paper presents an improved speaker verification technique that is especially appropriate for surveillance scenarios. The main idea is a meta-learning scheme aimed at improving fusion of low- and high-level speech information. While some existing systems fuse several classifier outputs, the proposed method uses a selective fusion scheme that takes into account conveying channel, speaking style and speaker stress as estimated on the test utterance. Moreover, we show that simultaneously employing multi-resolution versions of regular classifiers boosts fusion performance. The proposed selective fusion method aided by multi-resolution classifiers decreases error rate by 30% over ordinary fusion.

1 Introduction

Monitoring conversations in a communication network is a key tool in the counter-terrorism scenario. The goal may be either tracking known terrorists or even tracking suspicious conversations. In its basic form, trained personnel would listen to a sample of tapped conversations, sometimes guided by automatic word spot systems trained to detect suspicious vocabulary. Unfortunately, this solution is not feasible in practice for several reasons. First, the huge amount of simultaneous conversations precludes effective human surveillance. Moreover, criminals are often aware of tapped lines and communicate in codes, preventing the employment of word-spotting systems.

Automatic speaker recognition (ASR) and ‘stress’ detection systems could offer a practical solution at least towards reduction of the searching space. In theory, automatic systems could constantly sweep the network searching for enrolled criminals and suspiciously ‘stressed’ conversations. Relevant conversations could then be directed to human listeners. Current technology claims relative success both in the automatic speaker recognition field [1,2] and stress detection [3]. ASR technology has been assessed by NIST’s benchmarks [2] while detection of ‘stressed’ speech is still controversial, mostly due to the lack of proper databases for evaluations.

In this paper we present an improved speaker recognition system which takes into account the particular advantages and disadvantages of the surveillance scenario. The surveillance scenario is often characterized by long conversations that can be used as

training data for learning speaker models. On the other hand, unlike commercial applications in which the user will normally use his personal line number and handset, in the surveillance scenario a variety of different lines could be used, thus adding noise to the recognition process. Our method first automatically characterizes a conversation according to varieties of detectable noise, including speaker stress. It then exploits abundantly available training data to learn speaker models for speaker verification, where the particular model used is a function of the type of noise detected.

In particular, we refine recent work in speaker verification that exploits fusion of low and high speech levels classifiers [4]. These classifiers are based on a variety of feature types, including acoustic, phonetic, prosodic and even lexical ones. The method proposed by Campbell et al. [4] uses a linear combination of classifiers, employing a meta-learner to obtain optimal weights for the respective component learners.

In this work, we propose that the constituent learner weights not be assigned uniformly. Rather, the type and degree of distortion found in the speech sample to be classified is taken into account as part of the classification task. We show that by considering pre-defined data attributes, it is possible to fine-tune the fusion method to improve results. Thus, for example, although acoustic features are generally far superior to all other feature types, there are circumstances under which more weight should be given to lexical features. In a previous paper [5], we showed that a similar approach, in which the selective fusion is controlled by means of a decision tree, improves verification in simulated noisy conditions by more than 20%. In this paper, we exploit various types of “noise”, including channel characteristics and speakers’ emotional and stress patterns detectable in test conversations. Moreover, we show that including multi-resolution representations of some classifiers enhances fusion capabilities in handling noisy patterns, thus increasing accuracy.

This method both provides significantly improved speaker recognition accuracy as well as pinpointing ‘stressed’ conversations as a by-product. It is thus ideally suited for surveillance.

The organization of this paper is as follows. In section 2, speech production levels involved in the experiments and their implementation are presented. Experimental settings are presented in section 3. Sections 4 and 5 are dedicated to the proposed meta-learning scheme and results are presented in Section 6. In Section 7, multi-resolution classifiers are considered. Finally, conclusions and future research are discussed in Section 8.

2 Speech Levels

Humans can activate different levels of speech perception according to specific circumstances, by having certain processing layers compensate for others affected by noise. Utterance length, background noise, channel, speaker emotional state are some of the parameters that might dictate the form by which one will perform the recognition process. The present experiments seek to mimic this process. For this purpose, four classifiers were implemented targeting different abstract speech levels:

- The *acoustic* level, covered by a standard CEPSTRUM-GMM classifier. The term "acoustic" refers to the fact that the GMM spans the continuous acoustic space as defined by the CEPSTRUM features.
- The *phonetic* level, covered by a support vector machine (SVM) classifier using a feature set consisting of cluster indices provided by the GMM. We call this a "phonetic" classifier since it's based on counts of discrete acoustic units, namely, the GMM clusters. (To be sure, the term "phonetic" is not strictly appropriate, since we are not representing traditional phones, but rather abstract acoustic units resulting from clustering the CEPSTRUM space.)
- The *prosodic* level, covered by an SVM classifier using a feature set consisting of histogram bins of pitch and energy raw values and corresponding transitional tokens.
- The *idiolectal* level [6], covered by an SVM classifier using as a feature set frequencies of common words.

Generally speaking, the acoustic and phonetic levels are categorized as low-level as opposed to the higher prosodic and dialectal speech layers. Lower communication layers are normally constrained by the speakers' vocal-tract anatomy, while higher levels are more affected by behavioral markers.

Actually, the *acoustic* and *phonetic* levels are also represented in lower resolutions as analyzed in detail later in Section 7. In this case, less prototypical 'sounds' form the acoustic and phonetic space and each resolution is treated as a distinct (more abstract) level classifier. Let us now consider each of these in somewhat more detail.

2.1 GMM classifier

Our GMM implementation comprises a Universal Background Model (UBM) from which client models are derived through cluster mean adaptation and is very similar to that described in [1]. Only voiced frames are used. This decision was originally taken mainly in order to attain compatibility with the prosodic vectors stream. In this way, the vectors for all classifiers are obtained in parallel over the same time frames. The GMM consists of 512 gaussians, jointly trained for male and female speakers, taken from NIST'03 evaluation and no score normalizations (such as T- or Z-norm) [7] are performed. (Note that NIST'03 evaluation consists basically of cellular recordings, which are not ideal for modeling landline recording as in the present experiments. This is especially interesting in the context of this work, since we aim at investigating ways other non-acoustical sources compensate for GMM deficiencies. Moreover, unlike related work performed on this database [8]), no echo-canceling procedures were adopted in order to pre-clean this database.

2.2 SVM classifiers

Three separate SVM classifiers, one for each of the feature types – phonetic, prosodic and idiolectal – are implemented using the *SVMlight* package [9]. After some preliminary calibration, RBF was the chosen kernel for all SVMs with a radius of 10 for the phonetic and prosodic feature sets and a radius of 100 for the idiolectal feature set.

The phone vector is formed by accumulating the occurrences of the closest 5 (out of 512) GMM centroids for all utterance frames. Intuitively, this represents the speaker specific 'sounds set' frequency.

The prosody vector is formed by an agglutination of the following component counts:

- 50 histogram bins of the logarithmic pitch distribution;
- 50 histogram bins of the logarithmic energy distribution;
- 16 bi-grams of pitch-energy positive/negative time differentiates;
- 64 tri-grams of pitch-energy positive/negative time differentiates;

(There are four possible combinations for positive or negative pitch and energy slopes. Therefore, respectively, 4x4 (16) and 4x4x4 (64) possible bi/tri-gram tokens)

The idiolectal vector is formed by the entries of the 500 most frequency words found in the conversation transcripts.

Fusion of the four speech levels presented is implemented through extra linear SVM learners.

3 Experimental Settings

In this work, experiments are performed following the NIST'01 'extended data' evaluation protocol [10], based on the entire SWITCHBOARD-I [11] corpus. Only the 8-conversation training conditions were used. These comprise 272 unique speakers, 3813 target test conversations and 6564 impostor test conversations. Conversation lengths are 2-2.5 minutes. The evaluation protocol dictates a series of model/test matches to be performed. The matches are organized in 6 disjoint splits (we use splits 1, 2 and 3 for training and the others for testing), including matched and mismatched handset conditions and a small proportion of cross-gender trials. Besides speech files, automatic or manually generated transcripts are also available. In this work, we use BBN transcripts (available from NIST's site), which possess a word error rate of close to 50% (!) (Note that ordinary automatic transcripts can be easily obtained in surveillance applications.)

4 Data Attributes

A signal quality measure is needed as a means of controlling the fusion parameters, as a function of the degradation found in an utterance. We wish to use measurable attributes of the conversations to estimate the respective levels of three types of noise: communication channel, speaking style and speaker stress. Following is a brief description of the proposed attributes.

4.1 Channel

It is widely known that speaker recognition accuracy normally declines when the speech is conveyed through some communication channel. Real-world channels are band limited in frequency and often add some noise to the signal. Roughly speaking, an additive bias in acoustical features is introduced by different transmission lines. On the other hand, variance bias appears on the features due to additive (background) noise. Thus, means and standard deviations of the 20 filter bank outputs (byproduct of the MEL-CEPSTRUM extraction process) are retained as a representation of long-term channel behavior. In order to compress this information, which is highly redundant, DCT is applied to the means and the transformed first six components are kept. In addition, the mean of the individual filter bank standard deviations is calculated, as an estimation of additive noise level. One additional component will be added to the channel vector: the average GMM likelihood between the utterance frames and the background gaussian distribution (UBM). A low average likelihood will thus indicate an outlier feature distribution (possibly due to an unexpected channel).

4.2 Conversation Style

We approximate the emotional quality of an utterance through its pitch and energy averages and ranges, in addition to the estimated average articulation rate. In order to neutralize gender effects, pitch distributions are normalized per gender, so that male and female sets will possess the same mean pitch. Pitch and energy ranges are empirically measured as the number of histogram bins around the maximum value until a decay of, respectively, $1/4$ or $2/3$. Range asymmetry around the maximum is also included as the difference between the number of right and left bins. Articulation (or speech) rate is approximated as the average number of inflection points found in the first CEPSTRUM coefficient stream. A large average indicates a high rate of fluctuation of this parameter due to an accelerated speech rate.

4.3 Speaker Stress

Besides conversation style, we consider effects resulting from speaker stress. (Of course, the distinction between conversation style and speaker stress is somewhat arbitrary.) Features intended to identify stressed speech are based on the “Teager Energy Operator” (TEO) [3]. Our stress vector is composed of the mean and standard deviation of the TEO streams across 6 frequency bands.

5 Selective Fusion

In this section, we describe the proposed selective fusion method, which is depicted in Fig. 1.

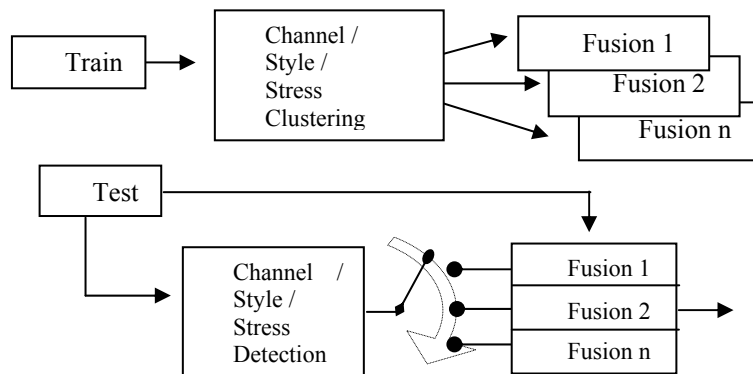


Fig. 1. Selective fusion method

In the training phase, k-means clustering is used to cluster the conversations according to respective attribute characteristics, namely channel, style and stress. (Note that only attributes and not explicit speaker verification features are employed in the selection phase.) Distinct fusion schemes are then learned for each cluster using linear support vector machines. In testing mode, each conversation is first assigned to the appropriate cluster according to its attribute profile and then the corresponding learned fusion scheme is applied.

Optimal ‘k’ (the number of clusters) and attribute vector composition (i.e. which of the attribute classes and components are the most efficient for data characterization) were selected through a nearly greedy search aiming at overall classification error reduction. Initially, full search was performed within the full three attribute classes separately. In a second step, the best candidates of each attribute class were concatenated in a composed vector and a new search was performed in order to determine optimal attribute vector composition. Clustering was performed on the basis of Euclidean distance after the vectors components were normalized to zero mean and unit standard deviation.

6 Results

In this section, we address the four different speech level classifiers, fusion results and the advantage obtained through the proposed meta-learning scheme. Results are analyzed also in light of the subjective ratings established by the SWITCHBOARD transcribers [11].

Individual classifier performance (in terms of percentage of misclassified conversation/speaker pairs, treating false positive and false negatives identically) and their relative weighting in ordinary fusion are depicted in Table 1. Fusion weights were obtained as the output of the fusion SVM, successively setting to 0 the scores corresponding to all individual classifiers except one. Bias terms were removed and the weights were then normalized to unity.

Table 1. Individual classifiers errors and weighting

	Acoust	Phon	Pros	Word
Error (%)	4.7	4.8	10.7	13.9
Weight	0.50	0.18	0.12	0.20

Table 2 summarizes fusion performance for the various attribute classes. Error rates shown are an average of 10 meta-learning runs (k-means clustering is non-deterministic).

Table 2. Fusion results

Attribute class	Error (%)
None (ordinary fusion)	2.84
Channel	2.63
Style	2.70
Stress	2.63
Channel + Style + Stress	2.39

As has been previously established [4], ordinary fusion offers better results than that obtained using any of the constituent speech levels individually. More significantly for our purposes, clustering and then fusing separately for each cluster offers improvement over ordinary fusion regardless of which attribute class is used. Maximal improvement is achieved, though, when all three attribute classes are considered. Let's focus on this case and consider a cluster scheme that proved to be optimal for our purposes. In Table 3 we show a stylized representation (in five quantization levels: {--, -, 0, +, ++}) for the attribute centroids derived from such clustering scheme. (Another successful cluster configuration included the 5th DCT channel component and the 2nd TEO band, instead of speech rate and energy asymmetry.) Recall that the optimal number of clusters ('k' in k-means) and attribute vector composition were found through greedy search.

Table 3. Fusion clusters

	speech rate	pitch range	energy asymm	3 rd TEO	4 th TEO
Cluster 1	+	++	+	+	+
Cluster 2	-	-	0	--	--
Cluster 3	+	-	++	++	++
Cluster 4	+	0	--	+	++

Table 4 shows individual classifiers weighting and verification error for each cluster.

Table 4. Fusion weighting and error

	Acoust	Phon	Pros	Word	Error (%)
Cluster 1	0.65	0.11	0.11	0.12	2.80
Cluster 2	0.38	0.28	0.19	0.16	0.87
Cluster 3	0.50	0.20	0.07	0.23	3.07
Cluster 4	0.48	0.24	0.03	0.25	3.53

Let us briefly analyze this fusion scheme configuration. Cluster 2 is the most accurate fusion set. According to subjective rating, it contains conversations with the smallest amount of echo, static and background noise. We have observed that conversations containing echo seem to be around 20% correlated with stress (TEO) values (see Table 3), as in this case. Moreover, absence of echo is associated with a decrease in the acoustic classifier share in the fusion process along with an increase in the phonetic classifier weighting. Possibly, the phonetic classifier, operating in a winner-takes-all fashion is quite sensitive to noise effects, since small perturbations may lead to erroneous ‘phone’ identification. On the other hand, likelihood values estimated by the acoustic classifier are more smoothly distorted. Actually, we will show in the next section that including acoustic and phonetic classifiers with a smaller number of phonetic units will improve fusion strategies in noisy environments.

In terms of speaking style, we observe a correlation between subjective ratings and attribute values. Cluster 1 is rated as the most natural sounding and bearing high topicality. In fact, the style components (see Table 3) possess high values for this cluster indicating vivid conversations. On the other hand, Cluster 2 is rated as relatively unnatural and bearing low topicality. Indeed, the low-valued style components for this cluster centroid indicate the presence of a formal speaking mode.

Similarly, one can concentrate only on stylistic or stress attributes in order to efficiently detect suspicious conversations in surveillance applications, although more profound analysis of the functions of these attributes remains to be done using stress/deception oriented databases. In fact, an auditory analysis of some stressed labeled utterances revealed that prominent low-stressed conversations (male only) sound extremely bass and as “newscast” style. On the other hand, the impression caused by very high-‘stressed’ conversations (female only) seems more difficult to typify. It seems that the high-‘stress’ does not reflect high pitch only.

7 Multi-Resolution

In this section, we show that simultaneous classifiers covering multi-resolution partitions of the low-level feature space highly boosts fusion accuracy. The motivation behind multi-resolution classification is to make available (a combination of) coarse and refined feature space clusterizations, which can be freely selected according to the nature of incoming test. We expect that noisy data would be more safely classified within a coarse segmented space, while clean data could explore the sharpness offered by a high-resolution mapping of the space.

For this purpose, we replicate the acoustic and phonetic SVM classifiers in 256, 128, 64 and 32-cluster resolutions, besides the original 512-cluster resolution. A greedy search was performed in order to find optimal ordinary fusion configurations. The following two configurations attained the lowest (**2.12%**) error rate:

- Acoust 512/256/64 + Phone 512/256/128 + Pros + Word
- Acoust 512/256/128/64 + Phone 512/256 + Pros + Word

Further error reduction can be achieved by applying selective instead of ordinary fusion. Optimal error reduction to **1.98%** was obtained for the former configuration. In this case, selective fusion is guided by two distinct attribute settings containing the 2nd and 6th TEO (stress) parameters and optionally one of the following: energy mean value or asymmetry. The following tables describe one of such settings. Table 5 schematically shows the attribute centroids for both clusters and the corresponding error rate for each fusion configuration. The unbalanced error rates are once more explained by the lesser amount of echo effects present in Cluster 1. This phenomenon is confirmed by the ratings assigned to the respective conversations and is in line with the low ‘stress’ assigned to Cluster 1.

Table 5. Stylized centroids

	Energy asymm	2nd TEO	6th TEO	Error (%)
Cluster 1	0	--	--	0.73
Cluster 2	0	+	+	2.72

Table 6 presents the weights assigned to each of the classifiers involved in the fusion process, for each cluster. It can be observed that for Cluster 2, the significance of low-resolution classifiers (Acoust 64 and Phone 128) is especially strong as compared to the corresponding higher-resolution versions. This is a reflection of the higher amount of noise in this cluster, requiring a decrease in feature-space resolution.

Table 6. Weights for individual classifiers

Classifier	Cluster 1	Cluster 2	Classifier	Cluster 1	Cluster 2
Acoust 512	0.20	0.64	Phone 256	0.10	-0.04
Acoust 256	0.20	0.31	Phone 128	-0.06	-0.16
Acoust 64	-0.02	-0.44	Pros	0.20	0.11
Phone 512	0.20	0.41	Word	0.18	0.19

8 Conclusions and Future Work

We presented in this paper a meta-learning scheme for fusion of several speech production levels. As opposed to standard classifier fusion, we introduce an utterance

quality measure, which adjusts the fusion scheme according to test signal idiosyncrasies. In addition, we show that multi-resolution low-level classifiers enhance fusion accuracy. Table 7 summarizes error reduction achieved with selective and multi-resolution fusion over the state-of-the-art GMM acoustic classifier and ordinary fusion of classifiers. Almost 60% error reduction could be achieved over the best individual classifier and 30% over ordinary fusion, with little calibration.

Table 7. Summary of fusion results

Classifiers configuration	Error (%)
Acoust (best individual classifier)	4.74
Ordinary fusion on: baseline (Acoust + Phone + Pros + Word)	2.84
Selective fusion on: baseline	2.39
Ordinary fusion on: baseline + Multi-resolution	2.12
Selective fusion on: baseline + Multi-resolution	1.98

The proposed scheme is well suited to surveillance applications. In this scenario, the presented sources of information can be easily extracted and are normally long enough to match the requirements for efficient fusion. In addition, the important function of detecting stressed conversation is already embedded in this scheme. Moreover, explicit stress detection can be achieved, pre-defining ‘stressed’ and ‘non-stressed’ conversation clusters within the selective fusion scheme, instead of unsupervised clustering through k-means. In this case, although optimum performance is not guaranteed anymore, suspiciously stressed conversations may be easily detected. Future work will focus on optimization of attribute characterization and selection and splitting of current speech features such as dynamic and static prosody and distinct dimensions of phonetic and acoustic representations. A deeper evaluation of stressed voiced detection is still pending the collection of appropriate databases.

References

1. Reynolds D., Quatieri T., Dunn R.: Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, Vol. 10, no. 1 (2000) 19–41
2. NIST - Speaker Recognition Evaluations: <http://www.nist.gov/speech/tests/spk/index.htm>
3. Zhou G., Hansen J.H.L., Kaiser J.F.: Nonlinear Feature Based Classification of Speech under Stress. *IEEE Transactions on Speech & Audio Processing*, Vol. 9, no. 2 (2001) 201-216
4. Campbell J., Reynolds D., Dunn R.: Fusing High- and Low-Level Features for Speaker Recognition. *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland (2003) 2665-2668
5. Solewicz Y. A., Koppel M.: Enhanced Fusion Methods for Speaker Verification. *9th International Conference “Speech and Computer” (SPECOM’04)*, St. Petersburg, Russia (2004) 388-392
6. Doddington G.: Speaker Recognition based on Idiolectal Differences between Speakers. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark (2001) 2517-2520

7. Auckenthaler R., Carey M., Lloyd-Thomas H.: Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, Vol. 10 (2000) 42–54
8. Andrews W. D., Kohler M. A., Campbell J. P., Godfrey J. J., Hernández-Cordero J.: Gender-Dependent Phonetic Refraction for Speaker Recognition. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Orlando, Florida (2002) 149-152
9. Joachims T.: Making large-Scale SVM Learning Practical. In: Schölkopf B. and Burges C., Smola A. (eds.): *Advances in Kernel Methods - Support Vector Learning*. MIT-Press (1999)
10. Przyboccki M., Martin A.: The NIST Year 2001 Speaker Recognition Evaluation Plan. <http://www.nist.gov/speech/tests/spk/2001/doc/> (2001)
11. SWITCHBOARD: A User's Manual. Linguistic Data Consortium, http://www ldc.upenn.edu/readme_files/switchboard.readme.html