

# Enhanced Fusion Methods for Speaker Verification

Yosef A. Solewicz (1,2) and Moshe Koppel (1)

(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 meta-learning schemes aimed at improving fusion of low and high level information for speaker verification in clean and noisy environments. While traditional systems fuse several classifier outputs in a uniform fashion independently of test quality, the proposed schemes use selective fusion weights according to test quality. A decrease of more than 20% under noisy conditions and 10% under clean conditions could be obtained with little calibration.

## 1. Introduction

In the problem of speaker verification, we wish to determine whether a given speech segment was spoken by a given speaker. Any potential solution must be rooted in the construction of models for the given speaker by comparing known speech of that speaker against some representative population. These models might be based on a variety of features including acoustic, phonetic, prosodic and even lexical ones. It has been shown that fusion of multiple such feature types yields more accurate models than can be obtained from any single feature type [1]. In this paper, we show that by considering possible discrepancies between training speech and a speech segment to be categorized, we can 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.

More specifically, we use Gaussian mixture models to exploit acoustic information and support vector machines to construct models from a variety of other feature types. Soft scores of each learned model are then used as input to a fusion model which itself is learned from labeled meta-examples. Stress discrepancies, channel discrepancies and other factors which might impinge on the reliability of one or another learned models can be measured using considerations of pitch, pitch difference and other measurable qualities of both training and test examples. In this paper, we focus on channel and noise issues inherent in test samples, which will degrade verification accuracy. In such cases, merging information is a known technique for attaining verification robustness. Previous approaches [1] blindly combine several classifiers through a meta-learner in

order to obtain more stable scores. On the other hand, we investigate novel meta-learning schemes in which the type and degree of distortion found in the speech sample to be classified is explicitly part of the classification task.

Our main result is that when these factors are used in the process of meta-learning fusion models, classification accuracy can be increased significantly in noisy conditions. Moreover, even clean test classification can be improved with the additional side information embedded in the learning scheme.

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. Finally, results and conclusions and future research are discussed in sections 6 and 7.

## 2. Fusion 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 which 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 frequencies of pitch-energy raw values and tokens.
- The *idiolectal* level [2], covered by an SVM classifier using as a feature set frequencies of common words.

Let us now consider each of these in somewhat more detail.

## 2.1. GMM classifier

Our GMM implementation comprises an Universal Background Model (UBM) from which client models are derived through cluster mean adaptation and is very similar to that described in [3]. Some minor differences are as follows. Our basic feature vector is composed by 19 MEL-CEPSTRUM coefficients, estimated in 40 msec windows, at a rate of 25 frames per second. (We use a quite large window, since we observed that it does not degrade performance.) Differentiates are appended to each vector by subtracting the consecutive and previous MEL-CEPSTRUM values. 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. Finally, mean and standard deviation normalization is applied to each vector stream. The GMM consists of 512 gaussians, jointly trained for male and female speakers, taken from NIST'03 evaluation. 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.

## 2.2. SVM classifiers

SVM classifiers are implemented using the *SVMlight* package [4]. 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 4 possible combinations for positive or negative pitch and energy slopes. Therefore, respectively  $4 \times 4$  (16) and  $4 \times 4 \times 4$  (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, are further fused as explained later through an extra linear kernel SVM learner.

## 3. Experiments settings

In this work, experiments are performed following the NIST 2001 'extended data' evaluation protocol [5], based on the entire SWITCHBOARD-I [6] corpus. Only the 16-conversation training conditions were performed, which are comprised of 57 unique speakers, 1328 target test conversations and 1368 impostor test conversations. Conversation lengths are 2-2.5 minutes. The evaluation protocol dictates a series of model/test matches which must be performed. The matches are organized in 6 disjoint splits, 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% (!).

In order to investigate fusion strategies under degraded conditions, a corrupted test set version is artificially created in addition to the original recordings. This is performed in two steps. Firstly, a random amount of pre-recorded 'crowd' noise is added to test recordings (mean SNR in noise  $\sim 14$  dB). Then the signal is band-pass filtered by an 8<sup>th</sup> order Chebyshev filter with random bandwidth (300-440 Khz to 2000-3400 Khz).

Training will also be optionally performed with pre-determined prototypical noise to cope with degenerated test data. Two levels ('Low' and 'High') of additive and bandwidth noise are defined to be equally located within the range of the random noise applied to the test patterns. Hence, the training set will be replicated into eight corrupted versions, apart from the original set. This results from the application of one or both noise simulations in two intensity levels. Thus, if we represent the original recordings by 'C' (clean) and denote 'L' and 'H' for low or high noise application in either additive or bandwidth modes, we have the following training set replications: {CC (original), CL, CH, LC, LL, LH, HC, HL, HH}. The degraded recordings will produce correspondent distorted feature streams. Nevertheless, we keep the word transcripts clean. Actually, this is a forced assumption, since we cannot reprocess the recordings with the original system responsible for

generating the transcripts. (We can partly justify this measure since, at least in theory, human transcriptions could be alternatively performed while current noise levels still retain text intelligibility.)

The four classifiers are separately trained in each of the nine defined noise conditions. Subsequently, classifiers' outputs are fused by extra SVM learners, producing fusion models for each of the nine categories. Splits 1 to 4 are used for training purposes. The test set consists of the original recordings ('clean') from splits 5 and 6 and their random noise version.

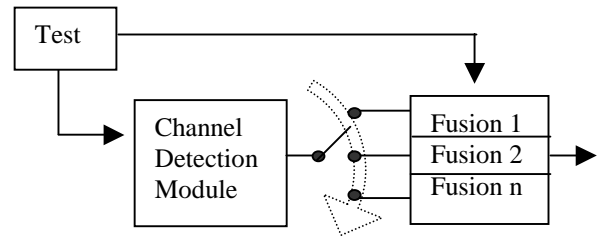
#### 4. Channel Characterization

One of the goals of this research is to evaluate fusion schemes in different test conditions. A signal quality measure is then needed as a means of controlling the fusion parameters, as a function of the degradation found in an utterance. We suppose that the degradations are mainly due to the channel conveying the speech signal. Therefore, we introduce a channel characterization vector, which will be a tentative measure of the amount and type of noise aggregated to the speech signal.

Our source of inspiration arises from channel normalization techniques. These techniques aim at applying some data normalization that will eliminate channel artifacts. 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 [7]. The simple way to combat these artifacts is by applying zero mean and unit variance normalization to the feature streams, a common procedure in speech recognition systems. Consequently, it can be assumed that features mean and variance are plausible candidates for a rough measure of signal quality. We then form the channel vector as follows. Means and standards deviations of the 20 filter bank outputs (byproduct of the MEL-CEPSTRUM extraction process) are retained. In order to compress this information, which is highly redundant, DCT is applied to the means and the transformed first six components are kept. They represent the basic shape of the transmission line. In addition, the mean of the individual filter bank variations is calculated. Once typical noise patterns are relatively spread over all frequency bands, it should be enough to keep an average value for the feature variations across the bands. One additional component will be added to the channel vector: the overall accumulated GMM distance between the utterance frames and the background gaussian distribution (UBM). The underlining idea is that if an utterance is extremely distorted, its features will not closely match the multi-modal gaussian distribution originally estimated to represent the non-distorted acoustic feature space. A large average distance will thus indicate an outlier feature distribution. In sum, the resulting channel vector will be composed of 8 (6+1+1) values.

#### 5. Meta-Learning

The general classification method employed is depicted in the following scheme:



“Figure 1. Meta-learning scheme”.

The basic idea is to address each test pattern to the most fitting fusion scheme learned, according to the test characteristics, through the ‘Channel Detection Module’. In this way, when noisy recordings are detected, predominant non-acoustical fusion systems would presumably be selected to classify this pattern, since they are known to be more noise resilient.

In this work the Channel Detection Module is realized either in a supervised or an unsupervised mode. The former is implemented through a Decision Tree (DT). The DT is trained with channel vectors derived from the original evaluation and its eight noisy versions (see session 3). Splits 1 and 2 are used for training and splits 3 and 4 are used for testing (and DT pruning). The DT is trained to classify each channel vector in one of the nine existing categories. The correspondent test pattern is then directed to the selected fusion scheme.

The unsupervised channel detection approach was implemented only for testing the uncorrupted original recordings. In this implementation, channel vectors are freely grouped through k-means clustering, each channel cluster being represented by a distinct fusion scheme.

The motivation behind this method is that one can consider test sub-categories even among the original SWITCHBOARD-I recordings. Possibly, discrimination between gender or handset (this corpus contains both carbon button and electret recordings) should be considered regarding fusion weights. Therefore, a conceivable evaluation might be allowing the channel vectors group by themselves in abstract prototype categories, in the hope that these categories would be better represented by distinct fusion models. (Obviously, a broader and optimized ‘channel vector’, which includes for instance pitch measures ought to be more appropriate in this case).

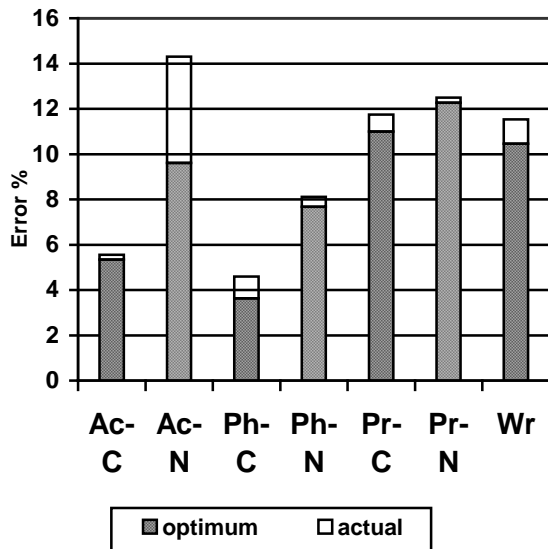
#### 6. Results

In this section, we present performance results in clean and noisy test conditions. We address the four different speech level classifiers, fusion results and the

advantage obtained through the proposed meta-learning scheme which directs tests to proper fusion schemes.

### 6.1. Individual classifiers

Figure 2 shows the error rates of the individual classifiers for splits 5 and 6 in clean (C) and noisy (N) conditions. Two types of errors (%) are depicted: the optimum, if the threshold is chosen ‘a posteriori’ and the correspondent actual error, obtained with thresholds learned from the original (clean) splits 1 to 4. (Recall that the word level does not possess a degenerated version.) It can be clearly seen that as expected the prosodic classifier is much more robust to noise than the acoustic GMM, which also shows a serious threshold shift in case the test is degraded.



“Figure 2. Individual classifier errors”.

### 6.2. Fusion results

Table 1 summarizes fusion performance under various conditions. The train quality field indicates the cases when only original recordings were used for training and those in which also noisy patterns were used. Correspondingly, results are shown separately for clean tests (that is, the original recordings) and for the artificially corrupted tests. Channel detection modes indicate cases in which the DT or k-means are used for switching among fusing classifiers as well as cases where a blind fusion of the training data is performed, with no discrimination between test patterns. The DT (random) mode reflects the error obtained after a random allocation of test patterns among the nine pre-defined noise conditions. It serves as a baseline for the regular DT performance.

In general, results suggest that in noisy conditions, simply fusing classifiers, even if noisy patterns are included in training, is not the best strategy. The preferable option is to apply selective fusion schemes, according to test pattern characteristics. Nevertheless, a

blind fusion still achieves better results for clean test conditions. This is because in this case, there is no need to train noise specific fusion schemes. Although the DT could simply direct all tests to the clean fusion scheme, DT allocation errors will direct clean tests to erroneous fusion classifiers. In fact, DT accuracy in classifying the nine types of noise-corrupted recordings (trained with the positive examples of splits 1 and 2 and tested in splits 3 and 4) was around 61%. When the DT was used for classifying the original recordings (of splits 5 and 6), only about 60% were labeled as ‘CC’ (no additive or bandwidth noises added).

The unsupervised channel detection showed surprising results. We performed a greedy search for optimal k (the number of clusters) and channel vector composition (i.e. which of the 8 vector channel components are the most efficient for channel characterization). It was observed (possibly due to limited amount of evaluation data) that some configurations seem to be quite sensitive to the k-means algorithm, since distinct runs lead to somewhat different errors. In general, for this database the most stable segmentation configuration was attained with the 3<sup>rd</sup> and 4<sup>th</sup> DCT components forming the ‘channel vector’ (reflecting secondary frequency effects) and two clusters splitting the channel space. The k-means scheme error shown in Table 1 was the one typically found for these settings, but in some runs even lower error rates were achieved.

“Table 1: Fusion results”.

Train quality	Channel detection	Test quality	Error (%)
Clean	No	Clean	2.13
Clean	No	Noisy	4.27
Clean +Noisy	No	Clean	4.48
Clean +Noisy	No	Noisy	3.76
Clean +Noisy	DT	Clean	2.56
Clean +Noisy	DT	Noisy	2.88
Clean +Noisy	DT(Random)	Clean	3.31
Clean +Noisy	DT(Random)	Noisy	3.84
Clean	K-means	Clean	1.92

### 6.3. Fusion weighting

It is interesting to gain some insight regarding the learned fusion schemes among the different speech layers in different conditions. The following table shows the weights fraction of each of the classifiers learned by the fusion SVM. As expected the fusion learned for noisy ('HH') training data strongly relies on the higher speech levels.

"Table 2: Normalized fusion weights per classifier".

	Acoust	Phones	Pros	Words
Clean	0.45	0.18	0.15	0.22
Noise	0.34	0.13	0.24	0.29

### 6.4. Unsupervised weighting

In one of the experiments we let the k-means algorithm find natural channel groups and then trained fusion schemes for the determined clusters. Table 3 depicts an example of weight distribution for two clusters. Recall that this method was applied only for the original (not artificially corrupted) dataset. We can see that even within clean recordings, optimal classification should emphasize different speech levels as a function of test quality.

"Table3: Normalized fusion weights per cluster".

	Acoust	Phones	Pros	Words
Fusion 1	0.45	0.21	0.06	0.28
Fusion 2	0.43	0.17	0.18	0.22

## 7. Conclusions and future work

We presented in this paper a meta-learning scheme for fusion of several speech production levels. As opposed to standard blind classifier fusion, we introduce an utterance quality measure, the 'channel vector', which controls the fusion scheme, according to test signal idiosyncrasies. This mechanism leads to about 23% decrease in error rate when dealing with artificially produced noisy tests, comparing to blind fusion. In addition, we showed that distinctive fusion schemes tailored for specific utterance classes are advantageous even in the case of clean tests. Initially, the utterance space is segmented through unsupervised clustering, based on their 'channel' characteristics. Finally, individual fusion schemes are trained for each class. This simple method yielded 10% improvement over plain fusion in the original evaluation.

Future work will focus on optimization of 'channel' components and inclusion of other sources of signal variability, such as speaker stress indicators. In addition,

other feature space segmentation techniques will be investigated.

## 8. References

- [1] Campbell, J., Reynolds, D., and Dunn, R. "Fusing high- and low-level features for speaker recognition", Proceedings of the 8<sup>th</sup> European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland, 2665-2668, 2003.
- [2] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," Proceedings of the 7<sup>th</sup> European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Denmark, 2517-2520, 2001
- [3] Reynolds, D., Quatieri, T., and Dunn, R., "Speaker verification using adapted gaussian mixture models". Digital Signal Processing 10, 1, 19-41, (2000)
- [4] Joachims T., "Making large-Scale SVM Learning Practical", Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- [5] Przybocki, M., and Martin A., "The NIST Year 2001 Speaker Recognition Evaluation Plan," <http://www.nist.gov/speech/tests/spk/2001/doc/>, March, 2001.
- [6] "SWITCHBOARD: A User's Manual," Linguistic Data Consortium, [http://www ldc.upenn.edu/readme\\_files/switchboard .readme.html](http://www ldc.upenn.edu/readme_files/switchboard .readme.html).
- [7] Viikki O. and Lurila, K., "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition", Speech Communication 25, 133-147 (1998).