Considering Speech Quality in Speaker Verification Fusion

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
koppel@netvision.net.il

Abstract

This paper emphasizes the benefits of embedding data categorization within fusion of classifiers for text-independent speaker verification. A selective fusion framework is presented which considers data idiosyncrasies by assigning particular test samples to appropriate fusion schemes. As an extension, incompatible data can be spotted and excluded from inherent classification errors. In addition, it's shown that multi-resolution low-level classifiers successfully boost fusion capabilities in noise.

1. Introduction

Recent work has shown great improvement in speaker verification accuracy through fusion of low and high speech levels classifiers [1]. 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. [1] 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 call it selective as opposed to ordinary fusion. We show that by considering pre-defined data attributes, including channel characteristics and speakers' emotional and stress patterns detectable in conversations, 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 this paper, strengths and weakness of feature sets are analyzed in light of subjective and objective speech quality as found in ordinary conversations. This represents an extension of a previous paper [2], in which a similar approach was successfully employed in simulated noisy conversations.

Moreover, we show that the inclusion of multi-resolution classifiers enhances fusion capabilities in handling noisy patterns, thus increasing accuracy.

Finally, we illustrate an added benefit of considering data characterization within the fusion scheme, namely, the possibility of rejecting atypical patterns before classification.

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. In Section 4, some objective data attributes are proposed and classifiers performance is analyzed according to subjective attributes. Selective fusion is presented in Section 5. In section 6, multi-resolution is considered and an example of outlier data removal is shown in Section 7. Conclusions are reported in Section 8.

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

2. Classification 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: acoustic, phonetic, prosodic and idiolectal [3]. Let us now consider each of these in somewhat more detail.

2.1. GMM classifier

Our GMM implementation targets the acoustic level and comprises a Universal Background Model (UBM) from which client models are derived through cluster mean adaptation and is very similar to that described in [4]. 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) 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. Moreover, unlike related work performed on this database [5], no echo-canceling procedures were adopted in order to pre-clean this quite contaminated database.) Although the acoustic classifier represents a relatively poor baseline, this is especially interesting in the context of this work, since we aim at investigating ways other non-acoustical sources compensate for GMM deficiencies in real unknown scenarios.

2.2. SVM classifiers

Three separate support vector machine (SVM) classifiers, one for each of the feature types – phonetic, prosodic and idiolectal – are implemented using the *SVMlight* package [6].

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 pitch and energy distributions and tokens of pitch-energy differentiates.

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 [7], based on the entire SWITCHBOARD-I 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%.

4. Data Attributes

Embedding data categorization in the fusion framework as a means of controlling the fusion parameters involves measuring the degradation found in an utterance. In this work, 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.

Channel characteristics are modeled by first and second statistics of the long-term spectrum of conversations. Means roughly reflect transmission line effects and variances point to additive noise level. The likelihood between an utterance frames and the UBM is used as an additional matching measure. Low likelihoods are expected for unseen channels, indicating that the specific conversation is not optimally modeled by the UBM.

Stylistic attributes are specified by means, ranges and symmetry of pitch and energy distributions. (Gender dependent mean pitch is first removed from all conversations.) In addition, an approximation of rate of speech is specified as the rate of fluctuations of the first Cepstrum coefficient. A large number of rises and falls indicate an accelerated rate of speech.

The Teager Energy Operator (TEO) [8] is a proposed measure of speaker 'stress'. We use first and second long-term statistics of TEO coefficients in six critical bands as an indication of this attribute. In fact, it was observed that this implementation is quite correlated with mean pitch and some normalization should be adopted in future experiments.

4.1. Rating Correlates

In order to illustrate the extent to which data quality can affect the different feature levels, we consider now a number of manually assigned characteristics of conversations and their impact on classification. As part of the transcription process, each SWITCHBOARD transcriber has to rate the conversations on a scale of 1 to 5, according to the following characteristics: (transcription) *Difficulty*, *Topicality*, *Naturalness*, *Echo*, *Static Noise*, *Background* (noise). We used these ratings in order to analyze classifiers performance in distinct situations. Only the most significant results, namely those obtained for *Topicality* and *Echo* are reported. For each of the rating categories, thresholds are obtained (from the first three splits) and applied on two partitions of the test set (the remaining three splits). One of the partitions contains all the conversations rated at the first level, respectively meaning: uniform topicality and low echo. The other test-set partition contains the remaining conversations ranked with higher grades, namely: multi-topicality and high echo. The groups represented around 40 to 60% of the whole test set. Table 1 shows error rates for each classifier as a function of the kind of test applied. ΔE represents the relative error increase (%), respectively from multi to uniform topicality and low to high echo.

Table 1: Performance according to "Topicality" and "Echo"

	Topicality			Echo		
Class.	Multi	Uni.	ΔE	Low	High	ΔE
Ac.	4.6	4.8	3.0	3.5	5.7	60.6
Ph.	5.1	4.7	-9.0	3.2	6.1	92.3
Pr.	9.7	11.1	14.0	10.1	11.0	8.4
Wr.	12.4	14.7	18.5	13.1	14.5	10.4

As might be expected, low-level classifiers are not directly affected by high-level speech issues such as the topicality level of conversations. Nevertheless, note that the high-level prosodic and word classifiers react better for conversations with varied topics. Possibly, some distinguishable intonational nuances are present during topic switching, enriching prosodic characterization. Moreover, high topicality expands vocabulary span in the test sample, which clearly improves idiolectal recognition capabilities.

Echo effects, on the other hand, are the strongest kind of degradation found in this database and responsible for the major part of the errors. Although a severe impact is felt on the low-level classifiers, echo damage is much less harmful to the high-level prosodic and word classifiers. (Actually, it's possible that the word transcripts used were obtained after echo canceling.)

5. Selective Fusion

In this section, we describe the proposed selective fusion method, which is depicted in Figure 1. 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 most effective for data characterization) were selected through a nearly greedy search aimed 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.

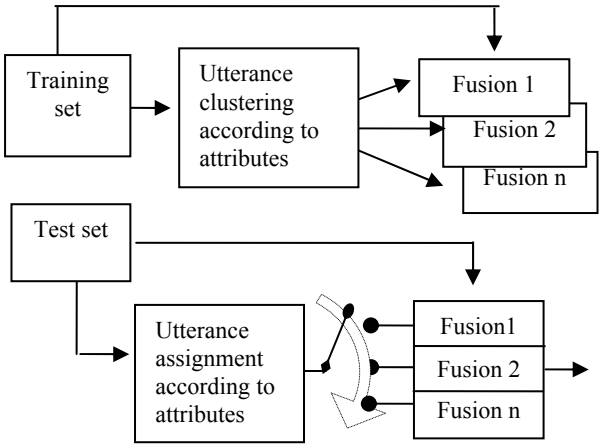

Figure 1: Scheme for selective fusion

Ordinary fusion of the four classifiers caused 2.84% of misclassified matches. On the other hand, within the proposed selective fusion, error dropped to 2.39% (estimated as an average of 10 meta-learning runs, since k-means clustering is non-deterministic).

Resulting fusion weights can be 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. Weights and error rates obtained through ordinary fusion of the individual classifiers are depicted in Table 2.

Table 2: Individual classifiers errors and weighting

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

Regarding selective fusion, two cluster schemes achieved the minimum error rate reported. In Table 3, we show a stylized representation (in five quantization levels: {--, -, 0, +, ++}) for the attribute centroids derived from one such clustering scheme. (The other 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=4 in this case) and attribute vector composition were found through greedy search.

Table 3: Fusion clusters and stylized centroids

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

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

Table 4: Fusion weighting and error

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

It is interesting to note that Cluster 2, which depends most heavily on phonetic features and on which we obtain highest accuracy, correlates strongly with those conversations which are subjectively classified as having lowest echo.

6. Multi-Resolution

In this section, we show that simultaneous classifiers covering multi-resolution partitions of the low-level feature space highly boost 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. Table 5 shows the ratio of weights between the highest and lowest resolution acoustic and phonetic classifiers and error rates for the two clusters formed. It's clear that the clean data cluster (bearing the lowest error rate) can more

efficiently exploit higher resolution resources, as opposed to the noisy cluster.

Table 5: High to low resolution weighting ratio and error

Cluster	Hi/Lo - Acoust	Hi/Lo - Phon	Error (%)
1	8.3	3.8	0.73
2	1.4	2.5	2.72

7. Rejecting Atypical Samples

In this section, we show how data attributes can be used to reject nonconforming samples prior to their classification. We aim at obtaining two partitions of the test set containing 'good' and 'bad' samples and reject the classification of the 'bad' partition since their components are likely to be poorly classified. Initially we define, within a development set, 'good' test samples as those obtaining (very) high scores for true matches and (very) low scores for impostor matches. Accordingly, tests that obtained scores close to the threshold area are labeled as 'bad'. A decision tree learning algorithm [9] using the attributes defined in Section 4 above is then used to distinguish 'good' and 'bad' test samples. In this database, the learned rule was simply to threshold the 6th TEO stress attribute, possibly due to its correlation with conversations containing echo. This means that continuous 'good' and 'bad' partitions of the test set could be achieved by means of a variable threshold sweeping the stress attribute scale. The moving threshold would successively define two complementary test clusters: those possessing stress higher than the threshold and those that do not.

Let's concentrate on the stress threshold leading to roughly even partitions over the test-set (splits 4, 5 and 6). We used a unique fusion configuration (**Acoust 512/256/64 + Phone 512/256/128 + Pros + Word**) trained on splits 1, 2 and 3, in order to classify the whole test-set and both its partitions. Now, we eliminate tests with scores close to threshold area in the three test groups (Figure 2). It can be observed that the error rate quickly decreases as poorly resolved tests (close to the threshold area) are discarded.

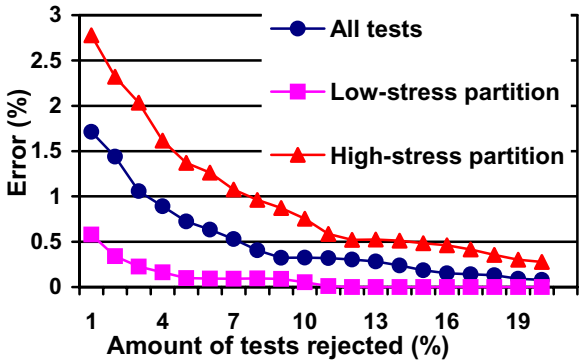

Figure 2: Decrease of error rate after rejecting bad tests

The ability to discern between non-conforming tests and inherent uncertainty is very important in applications such as forensics, since the inability to assess the strength of the evidence (due to noise, stress, etc.) doesn't necessarily mean that the prosecution hypothesis is untenable.

8. Conclusions

This paper investigated several aspects of the importance of considering data attributes within fusion of classifiers for speaker verification. We have shown that selective fusion in which fusion weights are optimized according to specific data attributes outperforms ordinary fusion. In addition, we have shown that the use of multi-resolution low-level classifiers enhances fusion capabilities. This feature could be particularly interesting when handling a variety of noisy patterns.

Future work should focus on refining data characterization, especially comprehending the link between objective and subjective attributes and further optimizing the selection of attribute vectors.

9. References

[1] Campbell J., Reynolds D., and Dunn R., "Fusing high and low-Level Features for Speaker Recognition", *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, pp. 2665-2668, 2003.

[2] Solewicz Y. A., and Koppel M., "Enhanced Fusion Methods for Speaker Verification", *Proceedings of the 9th International Conference "Speech and Computer"* (SPECOM'04), St. Petersburg, Russia, pp. 388-392, 2004.

[3] Doddington G., "Speaker Recognition based on Idiolectal Differences between Speakers", *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, pp. 2517-2520, 2001.

[4] Reynolds D., Quatieri T., and Dunn R., "Speaker Verification using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[5] Andrews W. D., Kohler M. A., Campbell J. P., Godfrey J., Hernández-Cordero, J., "Gender-Dependent Phonetic Refraction for Speaker Recognition". *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, pp. 149-152, 2002.

[6] Joachims T., "Making large-Scale SVM Learning Practical", *Advances in Kernel Methods - Support Vector Learning*. Schölkopf B., Burges C. and Smola A. (ed.), MIT-Press, 1999.

[7] Przybocki M., and Martin A., "The NIST Year 2001 Speaker Recognition Evaluation Plan", http://www.nist.gov/speech/tests/spk/2001/doc/, 2001.

[8] Zhou G., Hansen J.H.L., and Kaiser J.F., "Nonlinear Feature Based Classification of Speech under Stress", *IEEE Transactions on Speech & Audio Processing,* 9 (2): 201-216, 2001.

[9] Breiman L., Friedman J. H., Olshen R. A. and Stone C. J., *Classification and Regression Trees*, Wadsworth Int. Group, Belmont, California, 1984.