NOISE-ROBUST MULTI-STREAM FUSION FOR TEXT-INDEPENDENT SPEAKER AUTHENTICATION

Norman Poh and Samy Bengio

IDIAP, CP 592, 1920 Martigny, Switzerland

ABSTRACT

The the use of several acoustic feature types, also called the multi-stream approach, has proven to be very successful in speech recognition tasks and to a certain extent in speaker authentication tasks. In this study we propose a noise-robust multi-stream text-independent speaker authentication system. This system is trained in two steps: first train the stream experts under clean conditions and then train the combination mechanism to merge the scores of the stream experts under both clean and noisy conditions. The idea here is to take advantage of the rather predictable reliability and diversity of streams under different conditions. Hence, noise-robustness is *mainly* due to the combination mechanism. This two-step approach offers several practical advantages: the stream experts can be trained in parallel (e.g., by using several machines); heterogeneous types of features can be used and the resultant system can be robust to different noise types (wide bands or narrow bands) as compared to sub-streams. An important finding is that a trade-off is often necessary between the overall good performance under *all conditions* (clean and noisy) and good performance under *clean conditions*. To reconcile this trade-off, we propose to give more emphasis or prior to clean conditions, thus, resulting in a combination mechanism that does not deteriorate under clean conditions (as compared to the best stream) yet is robust to noisy conditions.

1. INTRODUCTION

Condition mismatch between training and testing is one of the most severe obstacles to making biometric authentication systems practical for day-to-day applications. This same problem is also encountered in speech recognition applications. Considerable amount of efforts have already been put to tackle the issue of robustness against mismatch conditions in the speech recognition community while not so in speaker authentication. The former community proposed the followings: multiple concatenated features (e.g. static and dynamic (more robust) features), the front-end multi-feature approach [1], multi-style training [2], the multi-stream approach [3] (whereby several features are used and processed independently), narrow-band training [4] (whereby several bands are used instead of several streams), Tandem [5], full combination approach [6] and more recently, the union model [7] (where several subsets of bands are combined and joint by the sum rule), to cite a few. In terms of theoretical concepts, however, both speech recognition and speaker authentication are almost identical. This study is somewhat inspired by many research works already done in the area of robust speech recognition through feature and model adaptation. However, there are several implementation differences that are unique to speaker authentication. The work done here hence will not only confirm findings in robust speech recognition but also demonstrate how similar concepts can be applied to speaker authentication. Therefore, emphasis will be put on the significant differences between both applications.

The focus of this paper is to deal with the robustness aspect of multi-stream fusion for speaker authentication. In a system with N speech feature streams (and hence N speech experts), and where some streams are more resistant to noise but deteriorate in performance in clean conditions while others perform better in clean conditions but deteriorate quickly in noisy conditions, we would like to know whether combining such streams at the score level will result in better performance under both clean and noisy conditions.

Section 2 outlines how multi-stream can be implemented for speaker authentication tasks. The proposed framework was tested on a NIST2001 database detailed in Section 3. A set of complementary features used are briefly explained in Section 4. This is followed by experimental results in Sections 5 and 6 (further experiments on controlling the priors), and conclusions in Section 7.

The authors thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on “Interactive Multimodal Information Management (IM2)”. The authors wish to thank Hynek Hermansky for giving constructing comments and Johnny Mariéthoz for providing many of the foundation works on which these experiments were carried out.

2. NOISE-ROBUST COMBINATION MECHANISM FOR MULTI-STREAM APPROACH

2.1. Classical and Multi-Stream Approaches to Noise-Robustness

Let us define real noisy conditions as $\mathcal{C}$, which consist of the tuple (noise type, SNR) where SNR is the corresponding Signal-to-Noise Ratio of the noise type. Note that both noise types and SNRs are not discrete. Hence $\mathcal{C}$ is continuous but often regarded as discrete in the literature. We will adopt such convention here. Let the utterance of a raw speech signal be represented by $\overrightarrow{X}$. A system (with parameters Θ) that is designed to be robust against mismatch conditions $\mathcal{C}$ can be represented by the following classical approach:

$$P\left(y|\{f(\overrightarrow{X_c})\}_{c\in\mathcal{C}},\Theta\right), \tag{1}$$

where f is a feature extraction function, such as Mel-scale Frequency Cepstrum Coefficients (MFCCs) and $y \in Y$ is a class label. For speech recognition, y could be class of phonemes (hence multi-class problem). For speaker authentication, y could be either client or impostor (two-class problem).

In the classical approach, the mismatch conditions are handled by incorporating such mismatch into the training phase of the system. This often results in deterioration of performance in clean conditions as well as in other *unseen* noisy conditions [2].

In the proposed multi-stream approach, this problem can be solved in two steps, as follows:

$$\overrightarrow{X_F^c} = \left[P(y|f_s(\overrightarrow{X_c}),\Theta_s)\right]_{s\in\mathcal{S}}, \text{ where } c = \text{clean}$$

and
$$P\left(y|\{\overrightarrow{X_F^c}\}_{c\in\mathcal{C}},\Theta_{COM}\right). \tag{2}$$

In the first step, one estimates the posterior probability of the class label y for each stream $s \in \mathcal{S}$ independently. Each stream-based system has the parameter set Θ_s. $\overrightarrow{X_F^c}$ is thus a vector whose elements are the hypothesis of stream $s \in \mathcal{S}$ that the feature vector $\overrightarrow{X_c}$ belongs to class y. Note that these stream-based systems are trained in clean conditions only. Noise-resistance is actually handled in step two, where each noisy condition $c \in \mathcal{C}$ is computed explicitly in the hypothesis space $\overrightarrow{X_F^c}$. Therefore, noise-robustness is due to this second step, which can be regarded as a fusion of $|\mathcal{S}|$ stream-based systems by a second classifier with parameter set Θ_{COM}. This classifier is called the COmbination Mechanism (COM). It should be emphasised that the multi-stream approach proposed in the literature (e.g. [6] in the context of speech recognition) often does not consider different noisy conditions when training the COM. The main idea here is to incorporate such noise-robustness into the COM.

In our opinion, the closest work to what is proposed here is by Cerisara et al [4] but in the context of speech recognition. They proposed to train the COM of a multi-band system (which is a Single-Layer Perceptron) in white noise, whereas the sub-band based experts are trained in clean conditions. The resulting multi-band system showed higher noise robustness to *most* of the noise cases tested, i.e., white noise, high-frequency (pink) noise, low-frequency (pink) noise, hair-dryer noise and car noise). Unfortunately, under clean conditions (and also in canteen noise), the performance of this system actually degraded compared to the system which is trained only in clean conditions. Hence, noisy conditions are tolerated with degraded performance in clean conditions.

In this paper, we opt for the multi-stream approach rather than for the multi-band approach. Both approaches are conceptually similar. The multi-stream approach exploits the use of different acoustic feature types. These feature types are selected such that the resultant feature-based expert have different performance in different conditions (see Section 4), i.e., some features perform better in clean conditions while others perform better in noisy conditions. This is the first major difference.

One possible advantage of using streams instead of using sub-bands is that the effect of coloured noise on sub-bands are unpredictable from one sub-band to another; whereas for streams, they are possibly predictable, i.e., one stream may be more robust to noise (coloured or not) than another. Since reliable streams/sub-bands should be weighted more than unreliable ones, the multi-band approach cannot exploit such prior knowledge while the multi-stream approach probably could. Hence, we expect that, due to the COM, the multi-stream approach will be robust to many kinds of noise types. This is also a hypothesis that we would like to validate. As far as we know, our proposed technique which makes use of such prior knowledge to fuse *opinions of stream-based experts* has not been applied neither for speech recognition nor for speaker authentication tasks.

2.2. Implementation Issues

In terms of implementation, we train the COM using "artificial noisy conditions". The motivation for doing so is that one does not know in advance how the real (noisy) conditions will be like. Hence, we propose to use an intuitive (although naive) class of artificial conditions, i.e., white noise at different SNRs. This practice has long been well-accepted in speech recognition community and has shown to work in [4]. Note that contrary to speech recognition, y in speaker authentication takes on two possible values: client or impostor. In speech recognition, step one is carried out using Gaussian Mixture Models (GMMs) or Multi-Layer Perceptrons (MLPs) (such as Tandem), whereas in speaker authentication, GMMs with Maximum *a Posteriori*

(MAP) adaptation have been the *de facto* approach because for practical reasons, they are found to be particularly suitable for such task. This difference is principally architectural.

A GMM models the statistical distribution of training feature vectors for each client. Briefly, a common impostor GMM model (also called a world model) is first obtained from many speakers using the Expectation-Maximization algorithm [8]. The world model is then adapted to each client's speech features using MAP estimation [9]. To make a decision, an average log-likelihood ratio between the client-adapted model and the world model (over all feature frames) is compared to a threshold chosen on development data.

As for step two (see Eqn. (2)), we propose to use MLPs and SVMs as the COM. This is because the output score vector $\overrightarrow{X}_F$ are highly correlated [10]. Non-linear mappings such as MLPs and SVMs provide a flexible means of finding the optimal separation hyper-plan. By simply analysing variance reduction due to averaging of $|\mathcal{S}|$ streams, it is known that the resultant combined system *cannot perform worse than the average performance* of $|\mathcal{S}|$ streams. Empirical evaluations in [10] showed that non-linear mapping *often* performs better than simple averaging, given that the right hyper-parameters (e.g. number of hidden units of MLP; standard deviation of SVM with Gaussian kernels) are used. They can be tuned by cross-validation. In fact, discrimination in streams is strongly desirable so that higher weights are given to the more reliable streams and vice-versa, under different conditions. The accept/reject decision is taken based on the output of the COM. This is detailed in Section 3.2.

We outline here a particular implementation of a noise-robust text-independent speaker authentication task in Algorithm 1. This algorithm takes in two data sets: training ($\mathcal{Z}_{\text{train}}$) and test ($\mathcal{Z}_{\text{test}}$) sets, which are taken from the *development* set. The first data set is used to train the base expert whereas the second set is used to train the COM. Step one consists of training the stream-based GMM experts $s = 1, \ldots, |S|$ independently using clean sequences of speech utterance.The number of Gaussians should be tuned by using the cross-validation technique. Step two consists of testing each s-th expert independently using both clean sequences and sequences corrupted by an artificially generated noise at different SNRs. Here, we use white noise at 18 dBs, 12 dBs, 6 dBs and 0 dBs. The procedure "train GMM" models the statistical distribution of $< f_s(\overrightarrow{X}), y >$ for y = client and y = impostor. The output of this procedure is a set of parameters describing the client and impostor distributions, i.e., Θ^s_C and $\Theta^s_{\overline{C}}$ respectively. The "test GMM" is a procedure that takes the client and impostor distributions, together with the test feature from a given noise condition c and outputs a LLR. This is done for each stream s. Hence, LLR^c is a vector of $|\mathcal{S}|$ elements, which is the number of

Algorithm 1 Robust multi-stream training ($\mathcal{Z}_{\text{train}}, \mathcal{Z}_{\text{test}}, \mathcal{S}, \mathcal{C}$)

$\mathcal{Z}$: pattern set $< \mathcal{X}, \mathcal{Y} >$
$\overrightarrow{X} \in \mathcal{X}$: training example
$y \in \mathcal{Y}$: the labels of training examples in $\mathcal{X}$
$s \in \mathcal{S}$: feature type (e.g. MFCC, LFCC)
$c \in \mathcal{C}$: condition
$\overrightarrow{X}_c \in \mathcal{X}_c$: example corrupted by condition c
STEP 1: train stream-based GMM experts
Use $(\overrightarrow{X}, y) \in \mathcal{Z}_{\text{train}}$
for each $s \in \mathcal{S}$ **do**
  $\{\Theta^s_C, \Theta^s_{\overline{C}}\}$ = train GMM ($< f_s(\overrightarrow{X}), y >$)
end for
STEP 2: test stream-based GMM experts under $\mathcal{C}$
Use $(\overrightarrow{X}, y) \in \mathcal{Z}_{\text{test}}$
for each $c \in \mathcal{C}$ **do**

$$\begin{aligned}\overrightarrow{X^c_F} &= \text{LLR}^c \\ &= \Big[\text{test GMM}\Big(\{\Theta^s_C, \Theta^s_{\overline{C}}\}, f_s(\overrightarrow{X}_c)\Big)\Big]_{s \in \mathcal{S}}\end{aligned}$$

end for
STEP 3: train the COM (MLPs or SVMs)
Θ_{COM} = train COM($< \bigcup_{c \in \mathcal{C}} \overrightarrow{X}^c_F, y >$)
return ($\forall_{s \in \mathcal{S}} \{\Theta^s_C, \Theta^s_{\overline{C}}\}, \Theta_{COM}$)

streams. Finally, step three consists of training the COM using the resultant set of vectors just mentioned. Here, we used MLPs or SVMs. The COM provides a mapping function from $\mathbb{R}^{|\mathcal{S}|}$ input dimensions to one output dimension where the final accept/reject decision will be taken once a threshold is determined. Our implementation of multi-stream text-independent speaker authentication is shown in Figure 1 with features described in Section 4.

This procedure has several advantages. Firstly, the subsystem s in $P(\overrightarrow{X}|\Theta_s)$ can be trained simultaneously (e.g., on different machines). Secondly, since the underlying streams can be trained and tested independently, it is therefore possible to use different window length, frame rates and different parameters to extract the features. This will be particularly useful to incorporate for instance time information at

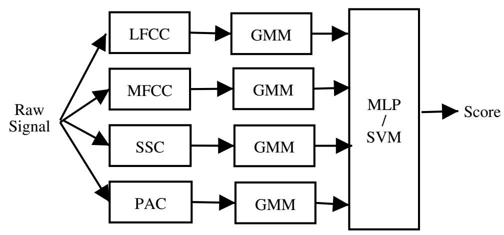

Fig. 1. A Multi-stream architecture

different modulation frequencies. Finally, only a subset of $\mathcal{C}$ will probably be needed to be considered by the COM, i.e., it does not have to take into account the artificial conditions at all the SNRs, because the reliability of streams are somewhat predictable. Hence, the COM will be able to generalise given proper tuning of hyper-parameters and regularisation.

3. EXPERIMENT SETUP

3.1. Database

The NIST2001 database [11] is used here. It is a commonly used (benchmark) database for text-independent speaker authentication tasks. The data is obtained from the Switchboard-2 Phase 3 Corpus collected by the Linguistic Data Consortium. Here, only the female subset (which is known to be slightly more difficult than the male subset) is used for evaluation. In the original database two different handsets were used (i.e., carbon and electret). However, only data from electret handsets are used (5 speakers who used the carbon handsets are removed) so that any variation of performance, if any, will not be attributed to this factor. This database was separated into three subsets: a training set for the world model, a development set and an evaluation set. The female world model was trained on 218 speakers for a total of 3 hours of speech. For both development and evaluation (female) clients, there was about 2 minutes of telephone speech used to train the models and each test access was less than 1 minute long. The development population consisted of 45 females while there were 506 females in the evaluation set. The total number of accesses for the development population was 2694 and 32029 for the evaluation population with a proportion of 10% of true accesses. Note that 4 types of noise: `white`, `oproom` (for operational room), `factory` and `lynx` noise, taken from the NOISEX-92 database [12], are used to contaminate the NIST2001 dataset.

3.2. Evaluation Criterion

In our multi-stream speaker authentication, the accept/reject decision is defined as:

$$\hat{F}(\overrightarrow{X}) = \begin{cases} accept & \text{if } p(y' = \text{client}|\overrightarrow{X_F}, \Theta_{COM}) > \Delta \\ reject & \text{otherwise} \end{cases} \tag{3}$$

Note that $\overrightarrow{X_F}$ derived in step one of Eqn. 2 is not used to make the decision. The decision is only made in step two as shown here. Because of this binary decision, the system may commit two types of error: false acceptance (FA) and false rejection (FR). FA happens when $\hat{F}(\overrightarrow{X}) = \text{accept}$ and $y = \text{impostor}$. FR happens when $\hat{F}(\overrightarrow{X}) = \text{reject}$ and $y = \text{client}$. They are quantified by False Acceptance Rate (FAR) and False Rejection Rate (FRR). Note that FAR and FRR are functions of the threshold Δ due to the fact that the decision function is itself a function of Δ. In this study, the commonly used Half Total Error Rate (HTER) is used as an evaluation criterion[1]. It is defined as $(\text{FAR}+\text{FRR})/2$. Here, we assume that the costs of false acceptance and false rejection are equal and that the prior (class) distribution of clients and impostors are equal as well. The HTER is calculated based on threshold Δ which itself is estimated *from a development set*. This threshold is estimated such that $|\text{FAR}(\Delta) - \text{FRR}(\Delta)|$ is minimised. It is then used to make decisions on an evaluation set. Hence, the HTER is *unbiased* with respect to the evaluation set since its associated threshold is estimated *a priori* on the development set. We call the resultant measure an *a priori* HTER and is used whenever an evaluation set is used. The lower HTER is, the better the performance.

4. MULTI-STREAM FEATURES

The main issue to examine here is: how to choose a good candidate feature set to be included into the multi-stream approach? According to the analysis of Variance Reduction (VR) [14], two systems should be as uncorrelated as possible. Hence, diversity is important. Since, the issue examined here is noise-robustness, this diversity should be with respect to noise-robustness, i.e., the candidate features should behave differently in different noise conditions. For instance, one feature should (result in experts that) perform well in one condition (e.g. clean) while another feature perform well in another condition (e.g. noisy condition at a given SNR). We have chosen four features that exhibit such complementary behaviour, listed in the order of decreasing accuracy in clean conditions (based on our experiments on the NIST2001 database), as follows:

- **LFCCs**: The Linear Filter-bank Cepstral Coefficient [15] speech features were computed with 24 linearly-spaced filters on each frame of Fourier coefficients sampled with a window length of 20 milliseconds and each window moved at a rate of 10 milliseconds. 16 DCT coefficients were computed to decorrelate the 24 coefficients (log of power spectrum) obtained from the linear filter-bank. (The same window length and shift were applied to all other features mentioned below.) The first temporal derivatives (deltas) were added to the feature set (so as other features described hereinafter). Furthermore, Cepstral Mean Subtraction (CMS) was also applied to normalise the effect due to channel distortion.

[1]It should be noted the popular Equal Error Rate (EER) *was not used* here because this criterion does not reflect real applications where a threshold must be fixed in advance. Moreover, the use of DET or ROC curves to compare two systems has recently been shown to be erroneous and misleading [13], despite the fact that they are widely accepted in the literature.

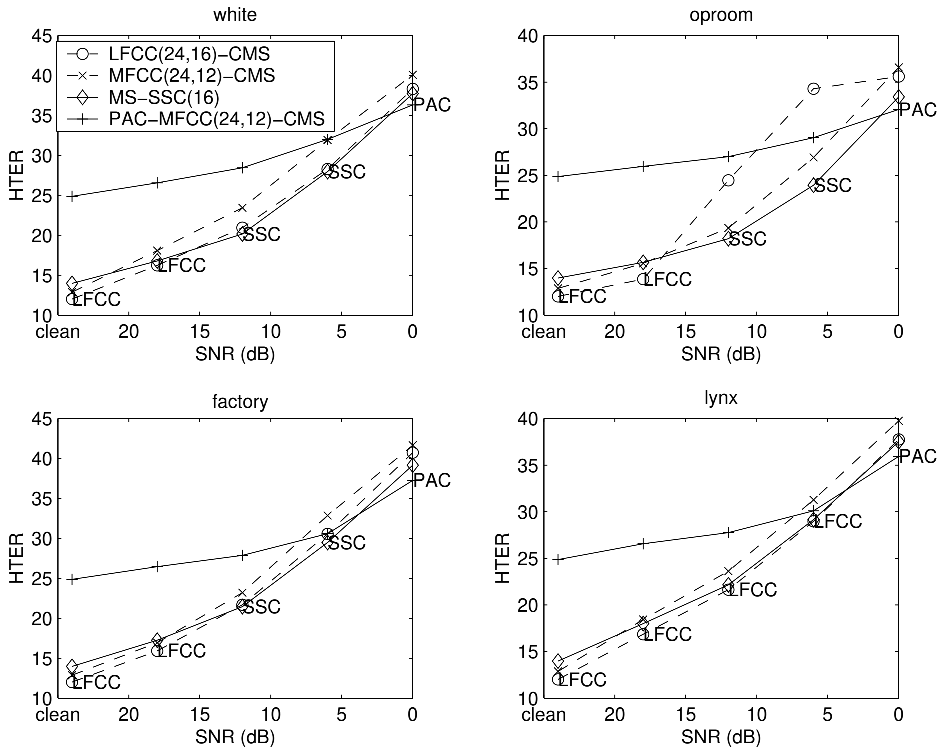

Fig. 2. A priori HTERs (in percentage; the lower the better) of various single-stream experts using GMMs under different noisy conditions, carried out on the female evaluation subset of the NIST2001 database. The labels are the best stream expert due to the associated feature set.

- **MFCCs**: The Mel-scale Filter-bank Cepstral. Coefficient [15] speech features were computed with 24 filters linearly spaced *on the Mel-scale* on each frame of Fourier coefficients. The first 12 DCT coefficients were computed to decorrelate the 24 coefficients (log of power spectrum) obtained from the filter-bank. Again, deltas were added and CMS was applied.
- **SSCs**: Spectral Subband Centroids [16, 17] are a set of centroids confined to be within each spectral subband. It was found that the mean-subtracted version of SSCs are more robust than the originally proposed SSCs. (Hereinafter, all SSCs imply mean-subtracted SSCs.) The SSCs used here were obtained from 16 centroids. The γ parameter, which is a parameter that raises the power spectrum, was set to 1.
- **PAC-MFCCs**: The Phase Auto-Correlation MFCCs [18] were extracted using 24 filter-banks spaced linearly on the Mel-scale, with 16 cepstrums. They are similar to MFCCs except that the the Fourier coefficients are derived uniquely from the phase-angle produced by auto-correlating speech waveforms instead of from both the magnitude and phase-angle of speech waveforms as commonly done. This has the effect of minimising the influence of additive noise. Experimental results in Section 5 show that PAC-MFCCs are very robust to noise but deteriorate greatly in clean conditions.

How would these features behave under different noise types and at different SNRs? A set of experiments is performed[2] and the results are shown in Figure 2.

The best stream in each noise type at a given SNR is labeled. It can be observed that LFCC features are the best set of features in clean conditions. Across different noisy conditions, it can be observed that SSC features turn out to be the best features while PAC features are the best in extremely noisy conditions. Note that there is a consistent behaviour across different noise types, thus, making such behaviour predictable in other noise types.

[2]The speech/silence segmentation is performed using two-competing Gaussians: one models the speech segments and the other models the silent segments. Note that we assume that this segmentation is uniform across different features and different conditions. Hence, it is performed once and applied to all other experiments. As a result, the experiments reported here are optimistically biased because in real situation, the segmentation of speech and silent cannot be reliably determined under noisy conditions. On the other hand, since the aim of this study is to measure the effectiveness of the COM against *noisy speech features*, it is undesirable that the unreliable segmentation (under noisy condition) be a variable factor that might make the analysis difficult.

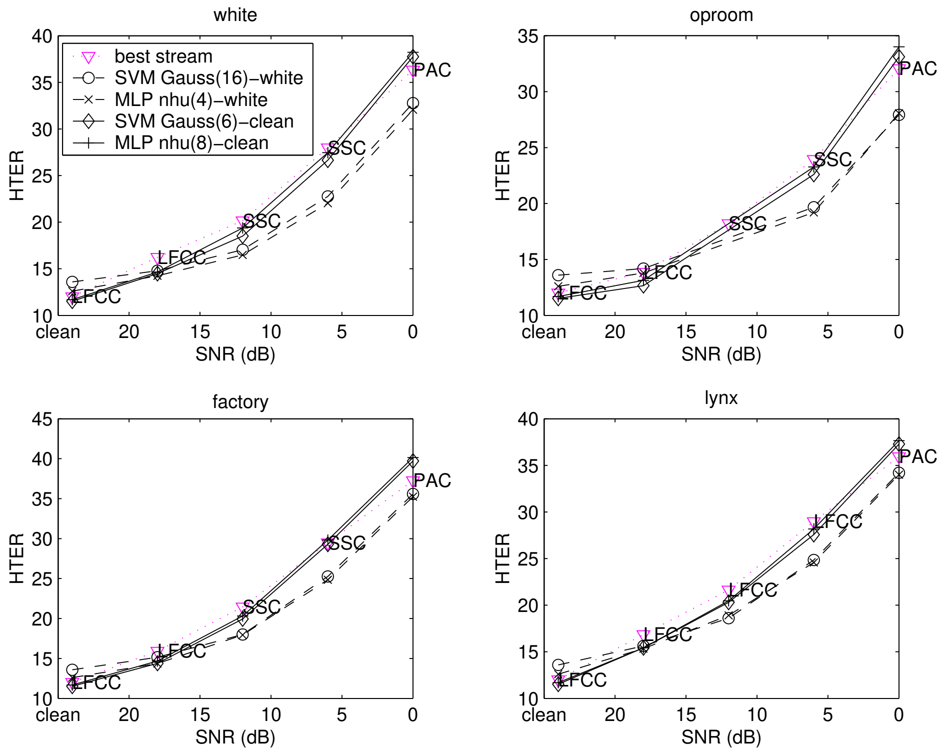

Fig. 3. A priori HTERs (in percentage) of various COMs (using MLPs and SVMs) trained under (1) clean conditions, and (2) under clean and various white noise conditions. The COMs were tested on the female evaluation subset of the NIST2001 database that were artificially corrupted by different noisy conditions. The COMs are labeled as [classifier type][hyper-parameters][conditions trained]

5. EXPERIMENTAL RESULTS IN MISMATCHED NOISY CONDITIONS

To test if prior knowledge of the COM is actually helpful or not, we first train the COM to fuse scores uniquely under clean conditions and the whole systems are tested on different conditions (different noise types at SNRs of 18, 12, 6 and 0 dBs). Both MLPs and SVMs are used as the COM. The second set of experiments consists of training the COM under clean and white noise at the mentioned SNRs. The generalisation performance of both sets of experiments under different noisy conditions *not seen during training* are shown in Figure 3. It can be observed that the former set of experiments performs well in clean conditions and gradually deteriorates (as compared to the best stream) under noisy conditions. In fact, under clean conditions, the MLPs and SVMs have a HTER of 11.684% and 11.518%, respectively and are significantly better than the best stream expert (i.e., LFCC expert, with HTER of 11.984%) according to the McNemar's test at 99% of confidence level[3] [19]. As for the latter set of experiments, the COM performances in fact deteriorate under clean conditions compared to the best stream expert (LFCC) but improve under noisy conditions. Under clean conditions, the MLPs achieved 13.096% of HTER while the SVMs achieved 14.428% of HTER. The best stream expert is better than both systems with 99% of confidence level according to the McNemar's Test. Clearly, the COM are trained to be be optimal in all conditions. Therefore, they lose to the best stream expert in clean conditions. It seems that there is a compromise to make: to excel in all conditions (hence losing in the clean conditions) or to excel in the clean conditions only (hence losing in noisy conditions).

6. CONTROLLING COMPROMISE BY PRIORS

Such compromise, in fact, can be expressed by viewing the problem in step two of Eqn. (2) *slightly* differently, i.e., c can be viewed as a target hidden state to be estimated. We

[3] This is done by calculating $((n_{01} - n_{10})^2 - 1)/(n_{01} + n_{10}) > p$ where p is the inverse function of $\mathcal{X}^2$ distribution (with 1 degree of freedom) at a desired confidence interval (i.e., 99%), and n_{01} and n_{10} are the number of *different* mistakes done by the two systems on the *same* accesses.

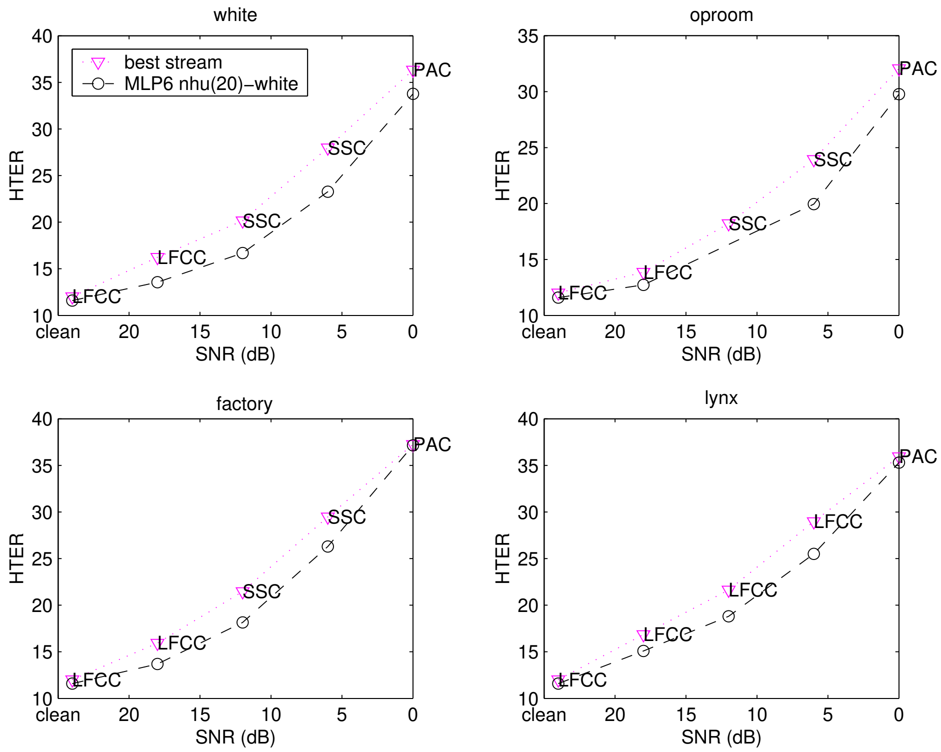

Fig. 4. A priori HTERs (in percentage) of the COM (using MLPs with 6 hidden units) trained under clean and various white noise conditions, with emphasis on clean examples. The COM was tested on the female evaluation subset of the NIST2001 database that were artificially corrupted by different noisy conditions.

propose to solve:

$$P(y, c|f(\overrightarrow{X_c}), \Theta) \tag{4}$$

Using Bayesian formulation, this can be solved using:

$$P(y, c|f(\overrightarrow{X_c}), \Theta) = \frac{1}{z}\underbrace{P(f(\overrightarrow{X_c})|\Theta, y, c)}\underbrace{P(y, c)}$$

where z is the normalising term $P(f(\overrightarrow{X_c})|\Theta)$. The first underbraced term is the conditional likelihood and the second underbraced term is the prior. Note that the prior $P(y, c) = P(y|c)P(c) = P(y)P(c)$ since y is independent of c.

From the two previous sets of experiments, it is found that indeed the COM can be better than the best stream expert under *clean conditions*. Therefore, to improve the COM under clean conditions and yet stay relatively robust under noisy conditions, one should give more prior to the clean conditions, i.e., $P(c = \text{clean})$.

There are several ways to incorporate such prior, which basically translates into giving more weights to examples (for both client or impostor class labels) in clean conditions. We chose the most straightforward way, i.e., by repeating N times the examples in clean conditions. This N is again controlled by cross-validation and is found to be 6. Hence, the proportion of the training data for the COM is 6:1:1:1:1 for the following conditions: clean, 18, 12, 6 and 0 dBs, respectively. To achieve similar effects in SVMs, one can control the soft-margin of each example, often represented as the C. It controls how much an example contribute to the margin. Unfortunately, by means of cross-validation on the C parameter[4] and the sigma parameter of Gaussian Kernel, we did not succeed in tuning these parameters to achieve the desirable output as done using MLPs. One can see the tuning of hyper-parameter as choosing one particular mapping function out of a set of infinite functions that will have an optimal performance under both clean and noisy conditions. This suggests that a specialised and more restrictive classifier would have been desirable rather than using a generic classifier with almost infinite capacities. The final results are shown in Figure 4. As can be observed, this COM gives a good trade-off in both clean and noisy conditions, across different noise types at different SNRs. Under clean conditions, the MLP achieves 12.163% of HTER while the best stream (LFCC) expert achieves 11.984% of HTER. Using McNemar's test, there is no significant difference at 99%

[4] We used suitable guess of C for different conditions, i.e., examples of the same noise condition are assigned similar value C, and each C value differs by an order of 10. Typical example of C for the conditions clean, 18, 12, 6 and 0 dBs are $\{100 : 10 : 10 : 10 : 10\}$.

of confidence level. Hence, with properly adjusted prior, a good performance trade-off of the COM in both clean and noisy conditions can be achieved.

7. CONCLUSIONS

We have proposed a noise-robust multi-stream text-independent speaker authentication system using a two-step approach: first train the stream experts under clean conditions and then train the combination mechanism (COM) to merge the scores of the stream experts under both clean and noisy conditions. The idea here is to take advantage of the rather predictable reliability and diversity of streams under different conditions. Hence, noise-robustness is due to the combination mechanism. This two-step approach offers several practical advantages: the stream experts can be trained in parallel (e.g., by using several machines); heterogeneous types of features can be used and the resultant system can be robust to all types of noise conditions. An important finding is that a trade-off is often necessary between the overall good performance under all conditions and good performance under clean conditions. To reconcile this trade-off, we proposed to give more emphasis or prior to clean conditions, thus, resulting in a combination mechanism that *did not deteriorate* under clean conditions (as compared to the best stream) yet stayed robust to noisy conditions.

Future studies in this direction will include analysing how the output hypothesis space is affected by the change in the feature space (which itself is affected by the raw audio signal) due to different Signal-to-Noise Ratios (SNRs). Ability to predict this change might give a hint on how to better combine the feature streams.

8. REFERENCES

[1] M. L. Shire, "Discriminant Training of Front-End and Acoustic Modeling Stages to Heterogeneous Acoustic Environments for Multi-Stream Automatic Speech Recognition," Ph.D. dissertation, University of California, Berkeley, USA, 2001.

[2] H.-G. Hirsh and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions," in *ISCA ITRW Workshop on Automatic Speech Recognition - Challenges for the New Millenium (ASRU2000)*, Paris, 2000.

[3] S. Dupont, "Étude et Dévelopment de Nouveaux Paradigmes pour la Reconnaissance Robuste de la Parole," Ph.D. dissertation, Laboratoire TCTS, Universite de Mons, Belgium, 2000.

[4] C. Cerisara, "Contribution de l'Approache Multi-Bande à la Reconnaissance Automatic de la Parole," Ph.D. dissertation, Institute Nationale Polytéchnique de Lorraine, Nancy, France, 1999.

[5] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, Istanbul, 2000, pp. 1635–1638.

[6] A. Hagen, "Robust Speech Recognition Based on Multi-Stream Processing," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne, Switzerland, 2001.

[7] J. Ming and F. J. Smith, "Speech Recognition with Unknown Partial Feature Corruption - a Review of the Union Model," *Computer Speech and Language*, vol. 17, pp. 287–305, 2003.

[8] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1999.

[9] D. A. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," vol. 10, no. 1–3, pp. 19–41, 2000.

[10] N. Poh and S. Bengio, "Non-Linear Variance Reduction Techniques in Biometric Authentication," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 123–130.

[11] A. Martin, "NIST Year 2001 Speaker Recognition Evaluation Plan," 2001.

[12] A. Varga and H. Steeneken, "Assessment for Automatic Speech Recognition: NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems." *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[13] S. Bengio and J. Mariéthoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication," Martigny, Switzerland, IDIAP Research Report 03-84, 2003, to appear in The Speaker and Language Recognition Workshop (Odyssey), Toledo, 2004.

[14] N. Poh and S. Bengio, "Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks?" Martigny, Switzerland, IDIAP Research Report 03-59, 2003, to appear in IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP), 2004.

[15] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Oxford University Press, 1993.

[16] N. Poh, C. Sanderson, and S. Bengio, "An Investigation of Spectral Subband Centroids For Speaker Authentication," Martigny, Switzerland, IDIAP Research Report 03-62, 2003, to appear in Int'l Conf. on Biometric Authentication, Hong Kong, 2004.

[17] K. K. Paliwal, "Spectral Subband Centroids Features for Speech Recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Seattle, 1998, pp. 617–620.

[18] S. Ikbal, H. Misra, and H. Bourlard, "Phase Auto-Correlation (PAC) derived Robust Speech Features," in *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, 2003.

[19] T. G. Dietterich, "Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.