TOWARDS SPEAKER AND ENVIRONMENTAL ROBUSTNESS IN ASR: THE HIWIRE PROJECT

A. Potamianos[1], *G. Bouselmi*[2], *D. Dimitriadis*[3], *D. Fohr*[2], *R. Gemello*[4], *I. Illina*[2], *F. Mana*[4], *P. Maragos*[3], *M. Matassoni*[5], *V. Pitsikalis*[3], *J. Ramírez*[6], *E. Sanchez-Soto*[1], *J. Segura*[6], *and P. Svaizer*[5]

[1] Dept. of E.C.E., Tech. Univ. of Crete, Chania, Greece
[2] Speech Group, "http://parole.loria.fr/" LORIA, Nancy, France
[3] School of E.C.E., Natl. Tech. Univ. of Athens, Athens, Greece
[4] Loquendo, via Valdellatorre, 4-10149, Torino, Italy
[5] ITC-irst, via Sommarive 18 - Povo (TN), Italy
[6] Dept. of Signal Theory, Networking and Communications, Univ. of Granada, Spain

ABSTRACT

In this paper, we present algorithms for dealing with variability and mismatch in speech recognition due to environmental conditions and non-native speaker populations. The proposed algorithms cover a broad spectrum of ideas including robust feature extraction, feature compensation and speech enhancement. Specifically the following algorithms are presented and evaluated: beamforming for multi-microphone speech recognition, robust modulation and fractal features, Teager energy cepstrum coefficients, parametric feature equalization, speech enhancement, and acoustic modeling for non-native speech recognition. Also the problem of feature fusion and voice activity detection are discussed. Evaluation results on the AURORA databases under the auspices of the HIWIRE project show that significant gains can be achieved under adverse or mismatched conditions using these algorithms. Relative error rate reduction of up to 50% was shown for multi-microphone speech recognition, robust feature combination and speech enhancement. 30-40% reduction was shown for parametric feature equalization and non-native acoustic models.

1. INTRODUCTION

Despite recent progress in the area of automatic speech recognition (ASR) performance in adverse conditions is still not satisfactory for many real-life application. In this paper, we present recent research achievements in the area of robust ASR dealing especially with adverse environments and speaker variability. A variety of algorithms for improving the performance of speech recognition system under adverse conditions is investigated, namely: incorporating additional information (multi-microphone processing), robust feature extraction, speech enhancement, feature equalization and voice-activity detection. In addition, algorithms for dealing with speaker variability, especially non-native speech are presented.

Distant-talking ASR is not yet a mature technology as several complications contribute to lower recognition performance when compared to an equivalent system operating with a close-talk input. Multi-microphone processing [4] can be used to obtain an enhanced version of the desired speech signal, by means of spatial filtering and selective acquisition of the speaker. In this paper, we present results of various beamforming approaches as a function of the number of microphones in the array.

Robust features can significantly improve the performance of a speech recognition system in noisy conditions. In this paper, three robust front-end features are proposed and evaluated. The proposed features are related to modulations phenomena in speech resonances, Teager-Kaiser energy measurements, dynamical systems and fractal theory. Some of these features are combined (fused) with traditional mel cepstrum coefficients (MFCC) using a multi-stream hidden Markov model (HMM) framework.

The problem of feature fusion in the context of multi-stream HMMs is an important problem. In the speech recognition literature, multi-stream recognizers have been used to combine feature streams of different reliability [18] or different information content [6, 24]. The problem of supervised stream weight computation is well studied: minimum error (discriminative) training can be used to select the best combination of stream weights during model training [23]. Recently there has been interest in investigating unsupervised algorithms for estimating stream weights during recognition [29]. In this paper, we present analytical results [22] for the selection of stream weights.

Feature normalization is a promising approach when mismatch exists between the conditions under which the acoustic model was trained and the conditions in the field. In this paper, an extension of the histogram equalization (HEQ) approach is proposed namely: parametric nonlinear equalization. Parametric equalization improves on HEQ by imposing constraints on the type of histogram transformation that can be applied between the testing and the training data.

Currently, there are technology barriers inhibiting speech recognition systems that work in extremely noisy conditions from meeting the demands of modern applications. These systems often require a noise reduction system in combination with a precise voice activity detector (VAD). In this paper, two novel VAD algorithms are presented one using bispectrum likelihood ratio test and another employing support vector machines for voice-silence classification. VAD algorithms can be combined with speech enhancement algorithms to further improve speech recognition accuracy and robustness to noisy conditions. In this paper, soft-decision gain modification for speech enhancement [5] is applied to speech recognition. The proposed method builds on the work in [8] and is evaluated on noisy speech recognition tasks.

This work was funded by the EU-IST HIWIRE project (http://www.hiwire.org/), the EU-IST MUSCLE project, and the Spanish MEC project TEC2004-03829/FEDER.

The drastic drop in performance for ASR systems when confronted with non-native speech is a well known problem. The main source of variability in non-native speech is pronunciation variation. In this paper, we improve ASR performance by incorporating prior knowledge and making the speech recognizer tolerant to pronunciation variants. Various approaches are used to extract this knowledge and integrate it into an existing ASR system [17].

The organization of the paper is as follows. In Section 2, algorithms related to environmental robustness in ASR are presented. Algorithms related to speaker variability and non-native speech are presented in Section 3. Experimental results for all the proposed algorithms are presented in Section 4 and the paper is concluded in Section 5.

2. ROBUSTNESS TO ENVIRONMENTAL CONDITIONS

In this section, we present algorithms for ASR that deal with adverse environmental conditions namely: beamforming for multi-microphone ASR, robust feature extraction, feature fusion, parametric feature equalization, voice activity detection and speech enhancement.

2.1. Multi-microphone Beamforming

A lot of literature has appeared in the last decades about techniques aiming at improving the quality of the desired speech and increasing the corresponding signal-to-noise ratio (SNR). The simplest - and effective - method is the "delay-and-sum beamforming", which is based on the temporal re-alignment of all the signals in order to compensate for the inter-channel delays due to the different distances between the desired source and the sensors. The effect of this approach is that any interfering signal is recombined out of phase and therefore attenuated.

One critical issue is the "time delay estimation" for the speech component in all the channels. It is the first step in many localization algorithms and turns out to be very critical for an accurate beamforming. The reference algorithm is based on the technique known as "Cross Power Spectrum Phase analysis" (CSP) [19] and relies on the detection of a maximum peak in the inverse Fourier transform of a normalized cross-spectrum between microphone pairs.

A method to speed up the delay estimation and to give further robustness exploits the assumption of fixed source (although in unknown position). This hypothesis suggests to enhance the estimation based on a single-frame basis by averaging the CSP over multiple frames. The idea derives from the observation that the sum of the DFT obtained from different frames, thanks to the linearity of the transform, is equivalent to a single DFT of a wrapped version $x_w(n)$ of the input signal $x(n)$ obtained by accumulation of the signal over the analysis window:

$$x_w(n) = \sum_{k=1}^{K} x(n + kL), \tag{1}$$

where k is the frame index and L is the number of points of the window. A single CSP computation, after having accumulated the two signals, is then sufficient to estimate the required delay.

Once obtained a set of mutual delays between the channels, a beamformed input is derived by the delay-and-sum method. We applied this procedure to assess its impact on recognition performance. A more sophisticated beamforming can produce higher SNR gain but at the expense of introducing a partial distortion of the desired speech component.

Another important issue regarding environmental acoustics and multi-microphone processing is the possibility to mimic in realistic way a given environment by means of proper measurements of characteristic parameters as impulse responses and background noises [16]. This allows in principle on one hand to predict the recognizer behavior without specific data and on the other hand to reduce the acoustic mismatch by training effective models tailored on the target environment. We adopted this framework to test the sensitivity of ASR performance to the environmental variabilities (see Section 4.1).

2.2. Robust Features

There has been strong experimental and theoretical evidence for the existence of important nonlinear aerodynamic phenomena in the vocal tract during speech production [15]; these indicate the existence of *modulations* and *turbulence* which may be generated during phonation. However, the state of the art in acoustic processing for ASR systems employs features like MFCCs that are based on the linear source-filter model. Further, even though several ASR systems have attained satisfactory performance, their efficacy degrades significantly when speech is contaminated with noise [13].

2.2.1. AM-FM Features

A speech resonance can be modeled by an AM-FM signal and the total speech signal as a superposition of a small number of such AM-FM signals [6]. Such a model suggests that the formant frequencies are not constant during a single pitch period. These variations are partly captured by the *Frequency Modulation Percentages* (FMP) features defined as $FMP_i = B_i/F_i$ for each speech resonance i; B_i is the mean bandwidth and F_i is the weighted mean frequency value of resonance i. Another frequency-related feature is the short-time weighted mean of the instantaneous frequency (IF) signal (IF-Mean) providing information about the speech formant fine structure. The fine structure of the amplitude envelope signal is measured with the mean of the *Instantaneous Amplitude* (IA) features (IA-Mean), as the short-time mean of the IA for each speech resonance.

2.2.2. TECC Features

The short-time average of the signal squared is widely used as an ad-hoc approximation of the energy of the signal's source. For resonance signals, the *Teager-Kaiser Energy* (TKE) provides a good estimation of the source energy. Herein, we employ a front-end that combines an auditory-motivated filter-bank with the TKE estimation method; these features are labeled *auditory Teager Energy Cepstrum Coefficients* (TECCs) [7] and their main differences with the MFCCs are the auditory filter-bank and the short-time TKE computation. In detail, we utilize a Gammatone filter-bank; its filters are smoother and broader than the ones of the MFCC triangular filter-bank, are denser in frequency and spaced according to the bark-scale. The logarithm of the short-time average of the TKE operator is computed for each band-passed signal and then the inverse discrete Fourier transform is applied to obtain the TECCs.

2.2.3. Fractal Features

The fractal dimension can be interpreted as an approximate quantitative characteristic feature that corresponds to the amount of turbulence that may reside in a speech waveform. Extending previous work [15], we present a *combination* of dynamical filtering

on embedded noisy speech signals followed by correlation dimension measurements (Filtered Dynamics – Correlation Dimension, FDCD). In addition we incorporate the multi-scale fractal dimension (MFD) [15] in all experiments. The embedding vector defines a motion in a reconstructed multi-dimensional space; if the unfolding is successful the resulting system has common invariants with the original one [21]. In the *unfolded* phase-space we measure the correlation dimension (CD) and form a feature vector comprising its statistics [20]. Prior to the CD measurements we employ a denoising method in the unfolded phase-space [21] by iteratively decomposing the local neighborhoods to a set of eigenvectors and projecting on the subspace spanned by the largest principal components.

2.3. Feature Fusion

A common practice for combining information sources in a statistical classification framework is the use of "feature streams". In this section, we investigate the problem of unsupervised stream weights computation. Analytical results for the selection of stream weights as a function of single-stream estimation and misclassification errors for the two class problem are presented next. Two cases are investigated (see [22] for details):

- **Equal Bayes classification error**: We assume that each of the single-stream classifiers have the same Bayes classification error but different estimation errors. In this case, we also make the assumption that in the decision region $p(x_1|w_1) \approx p(x_2|w_1)$, provided that the features x_1, x_2 follow a similar parametric distribution (e.g., Gaussian) for the two classes w_1, w_2, and are variance-normalized. The weights s_j that minimize the estimation error are given by:

$$\frac{s_1}{s_2} = \frac{\sum_{i=1}^{2} \sigma_{i,2}^2}{\sum_{i=1}^{2} \sigma_{i,1}^2} = \frac{\sigma_{S2}^2}{\sigma_{S1}^2} \tag{2}$$

  where σ_{S1}^2 and σ_{S2}^2 are the single stream estimation variance, i.e., *the stream weights are inversely proportional to the variance of the PDF estimation error for each stream*. If the PDF estimation error variance in the two stream is equal then stream weights are equal, i.e., no stream weights should be used.

- **Equal PDF estimation error variance**: We assume that the (stand-alone) single-stream classifiers have the same PDF estimation error variance, but different classification errors, i.e., $\sigma_{S1} = \sigma_{S2}$. In this case

$$\frac{s_1}{s_2} \approx \frac{p(x_2|w_1)}{p(x_1|w_1)} \quad \text{for} \quad 0.5 \leq \frac{p(x_1|w_1)}{p(x_2|w_1)} \leq 1.5 \tag{3}$$

  i.e., in the region of interest *the stream weights should be inversely proportional to the classification error of the single-stream classifiers*. Note that if $p(x_2|w_1)/p(x_1|w_1) \geq 2.72$ the estimation error is minimized by setting one of the two stream weights to zero, i.e., if $p(x_2|w_1) >> p(x_1|w_1))$ then $s_1 = 1$ and $s_2 = 0$.

From the equations above it is easy to see that stream weights may reduce estimation error only when either the PDF estimation errors of the single-stream (stand-alone) classifiers are different, i.e., *one feature stream is more reliable than the rest*, and/or the Bayes errors of the single-stream classifiers are different, i.e., *one stream contains more information pertinent to the classification problem than the rest*. These results agree with our intuition and the results from experiments using supervised discriminative algorithm for estimating stream weights.

The theoretical results presented here can be used to obtain estimates of single-stream classification error from test data to address the unsupervised stream weight estimation problem. More work is underway to help us better understand the applicability of the optimal stream weight results to multi-stream recognition using HMM models.

2.4. Parametric Nonlinear Equalization

A new front-end normalization algorithm that uses a parametric nonlinear transformation of the voice features has been proposed and implemented [9]. This method improves the histogram equalization technique (HEQ) [30], by finding a simple and computationally inexpensive parametric expression of the nonlinear transformation done by HEQ. This new parametric approach relies on a two Gaussian model for the probability distribution of the features: a Gaussian model for the speech frames and a Gaussian model for the non-speech frames. A simple Gaussian classifier is used to label the input frames as belonging to one or the other class. For each class, the parametric linear transformation is defined to map the clean and noisy representation spaces, as described in the equations bellow:

$$\begin{aligned} \hat{x} &= \mu_{n,x} + (y - \mu_{n,y}) \left(\frac{\Sigma_{n,x}}{\Sigma_{n,y}} \right)^{1/2} \quad \text{if } y \text{ is non-speech} \\ \hat{x} &= \mu_{s,x} + (y - \mu_{s,y}) \left(\frac{\Sigma_{s,x}}{\Sigma_{s,y}} \right)^{1/2} \quad \text{if } y \text{ is speech} \end{aligned} \tag{4}$$

where $\mu_{n,x}$, $\Sigma_{n,x}$, $\mu_{s,x}$ and $\Sigma_{s,x}$ correspond to the Gaussians modeling clean non-speech and speech frames, respectively, and $\mu_{n,y}$, $\Sigma_{n,y}$, $\mu_{s,y}$ and $\Sigma_{s,y}$ correspond to the Gaussians modeling noisy non-speech and speech frames being equalized. With these definitions of the linear transformations, the noisy means $\mu_{n,y}$ and $\mu_{s,y}$ are transformed into the clean means $\mu_{n,x}$ and $\mu_{s,x}$, and the noisy covariance matrices $\Sigma_{n,y}$ and $\Sigma_{s,y}$ are transformed into the clean covariance matrices $\Sigma_{n,x}$ and $\Sigma_{s,x}$ (for both, the non-speech and speech models). The clean Gaussians for speech and non-speech frames can be estimated from the training database, while the noisy Gaussians should be estimated from the utterance to be equalized.

The result is a more robust equalization that improves 2 drawbacks of histogram equalization. The first drawback faced, is the fact that in most cases HEQ is based on estimations that count on a reduced number of observations belonging to the utterance in process of equalization. Using the parametric models, with little free parameters a smoother estimation is achieved. The second drawback faced is the dependence on the amount of non-speech frames of each utterance when estimating the CDF. The variable number of non-speech frames introduces an unwanted variability in the estimated CDF. The proposed algorithm creates separate modes for non-speech frames and speech frames eliminating with this the undesired random variation.

2.5. Voice Activity Detection

This section summarizes the different methods that are being analyzed for robust VAD.

2.5.1. Integrated bispectrum likelihood ratio tests

One of the most important disadvantages of VAD methods based on power spectrum divergence measures is that no *a priori* information about the statistical properties of the signals is used. Higher order

statistics methods rely on an *a priori* knowledge of the input processes and have been considered for VAD since they can distinguish between Gaussian signals (which has a vanishing bispectrum) from non-Gaussian signals. However, the main limitations of bispectrum-based techniques are that they are computationally expensive and the variance of the bispectrum estimators is much higher than that of power spectral estimators for identical data record size. We have developed different approaches for effective VAD based on contextual likelihood ratio tests defined on the integrated bispectrum of the noisy speech that has reported significant benefits in robust speech recognition applications [12, 25]. It inherits the ability of higher order statistics to detect signals in noise with many other additional advantages: *i*) its computation as a cross spectrum leads to significant computational savings, and *ii*) the variance of the estimator is of the same order as that of the power spectrum estimator.

The problem is then formulated in terms of a classical binary hypothesis testing framework. Given an observation vector $\hat{\mathbf{y}}$ to be classified, the problem is reduced to selecting the class (H_0 or H_1) with the largest posterior probability $P(H_i|\hat{\mathbf{y}})$. Thus, a statistical LRT is defined as:

$$L(\hat{\mathbf{y}}) = \frac{p_{\mathbf{y}|H_1}(\hat{\mathbf{y}}|H_1)}{p_{\mathbf{y}|H_0}(\hat{\mathbf{y}}|H_0)} \tag{5}$$

and the observation vector $\hat{\mathbf{y}}$ is classified as H_1 if $L(\hat{\mathbf{y}})$ is greater than $P(H_0)/P(H_1)$ otherwise it is classified as H_0. Assuming the integrated bispectrum $\{S_{yx}(\omega) : \omega\}$ as the feature vector $\hat{\mathbf{y}}$ and to be independent zero-mean Gaussian variables in presence and absence of speech, the evaluation of the test only requires to estimate the integrated bispectrum of the noisy signal and its variance. A careful evaluation of the proposed method [12, 25] shows clear improvements in detection accuracy and speech recognition over standardized VADs and over a representative set of recently published VAD algorithms.

2.5.2. Support vector machines

Since their introduction in the late seventies, support vector machines (SVMs) marked the beginning of a new era in the learning from examples paradigm. Detecting the presence of speech in a noisy signal is a two-class classification problem requiring a rule, which, based on external observations, assigns an object to one of the classes. A possible formalization of this task is by means of SVMs that enable building a function $f : R^N \longrightarrow \{\pm 1\}$ using training data that is, *N*-dimensional patterns $\mathbf{x}_i$ and class labels y_i. SVM enables to redefine the classification problem into some other potentially much higher dimensional feature space via a nonlinear transformation Φ:

$$f(\mathbf{x}) = sign\{\sum_{i=1}^{\ell} \nu_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b\} \tag{6}$$

where the dot product is efficiently computed according to the Mercer's theorem by means of kernels defined to be $k(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$, and the weights ν_i are the solution of a dual optimization problem. Once the SVM model is trained, the speech features **x** consisting on a contextual representation of wideband SNRs are classified according to the SVM decision function [26].

2.6. Speech Enhancement

A soft-decision gain modification for speech enhancement (but not for speech recognition) has been proposed in [5]. In the proposed method, a different soft decision gain modification is introduced and applied to the Ephraim-Malah gain function based on Maximum Mean Square Error Estimation (MMSE) [8] after amplitude compression. Non-linear evaluations of the noise overestimation factor and spectral floor are used in the same way for the proposed gain modification and for non-linear Spectra Subtraction (NSS) with Wiener filter. Consistent and statistically significant ASR improvements of the proposed approach with respect to NSS are observed for different noise conditions in the Aurora-3 corpus. As the non-linearity affects the two approaches in the same way, the result of comparison is particularly interesting.

Let $|Y_k(m)|^2$ be the k-th frequency sample of the spectrum energy of the noisy signal Y, computed in the m-th time window. Let $|X_k(m)|^2$ and $|D_k(m)|^2$ be the k-th spectrum energy sample, computed in the m-th time window, of the clean signal and the additive noise, respectively. $|X_k(m)|^2$ can be estimated using a Wiener filter, whose transfer function is $G_k(m)$, to compute:

$$|X_k(m)|^2 = G_k(m)|Y_k(m)|^2. \tag{7}$$

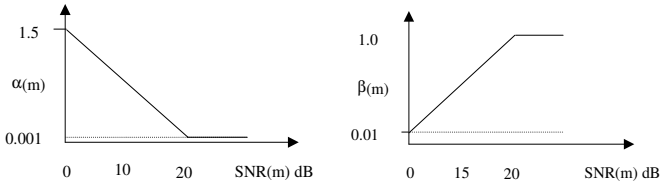
In [28], it has been found that good results are obtained if the filter is used to perform a non-linear spectral subtraction to compute $|X_k(m)|^2$ as follows:

$$\begin{cases} \frac{[|Y_k(m)|^2 - \alpha(m)|D_k(m)|^2]^2}{|Y_k(m)|^2} & \text{if } |Y_k|^2 - \alpha|D_k|^2 > \beta \\ \beta(m)|Y_k(m)|^2 & \text{otherwise} \end{cases} \tag{8}$$

where $\alpha(m)$ is a noise overestimation factor, and $\beta(m)$ is a spectral floor used to avoid negative spectrum values. These two parameters vary in time as function of the Signal-to-Noise Ratio $SNR(m)$, computed as follows:

$$SNR(m) = 10\ log_{10}\left(\frac{\sum_k |Y_k(m)|^2}{\sum_k |\hat{D}_k(m)|^2}\right) \tag{9}$$

where $|\hat{D}_k(m)|^2$ is an estimation of the k-th noise spectral sample at time m; $\alpha(m)$ and $\beta(m)$ are defined as possibility functions of $SNR(m)$ as shown next:



$G_k(m)$ can also be obtained with an approach proposed in [8]. In particular Ephraim-Malah MMSE log estimator is a short-time spectral amplitude estimator that minimizes the mean-square error of the estimated logarithms of the spectra, and it is well known that a distortion measure which operates on these logarithms is more suitable for speech processing than measures taken on the power spectra. It is defined as follows:

$$G_k = \frac{\xi_k(m)}{1 + \xi_k(m)} \exp\{\frac{1}{2}\int_{v_k(m)}^{\infty} \frac{e^{-t}}{t} dt\} \tag{10}$$

where:

$$\xi_k(m) = \frac{|X_k(m)|^2}{|D_k(m)|^2}, \quad \gamma_k(m) = \frac{|Y_k(m)|^2}{|D_k(m)|^2} \tag{11}$$

$\xi_k(m)$ is the *a priori* SNR, and $\gamma_k(m)$ is the *a posteriori* SNR. Also $v_k(m) = [\xi_k(m)/(1 + \xi_k(m))]\ \gamma(k)$.

The computation of the *a priori* SNR requires the knowledge of the clean speech spectrum, which is not available. An estimation can be obtained with a *decision-directed approach* as follows:

$$\hat{\xi}_k(m) = \eta(m)\frac{|\hat{X}_k(m-1)|^2}{|\hat{D}_k(m-1)|^2} + (1-\eta(m))\max\{0, \gamma_k(m)-1\} \quad (12)$$

In [1], it is shown that it is convenient for speech coding to make $\eta(m)$ dependent on the global $SNR(m)$ and to assign to it a high value if $SNR(m)$ is low and a low value if $SNR(m)$ is high. This method proposes to make the estimation of the *a priori* and the *a posteriori* SNR dependent on the noise overestimation factor $\alpha(m)$ and the spectral floor $\beta(m)$ as follows:

$$\hat{\xi'}_k(m) = \frac{\eta(m)|\hat{X}_k(m-1)|^2}{\alpha(m)|\hat{D}_k(m-1)|^2} + (1-\eta(m))(\gamma'_k(m)-1), \beta(m)\} \quad (13)$$

and

$$\gamma'_k(m) = \max\{\frac{|\hat{Y}_k(m)|^2}{\alpha(m)|\hat{D}_k(m)|^2} - 1, \beta(m)\} + 1 \quad (14)$$

where the noise overestimation factor $\alpha(m)$ and the spectral floor $\beta(m)$ varies with $SNR(m)$ as shown above. The adopted approach modifies the estimates of γ_k and ξ_k while maintaining the global shape of the gain function $G_k(\gamma_k, \xi_k)$. The modified gain function can be expressed as $G'_k(\gamma_k, \xi_k) = G_k(\gamma'_k, \xi'_k)$ with γ'_k, ξ'_k computed according to (13), (14). Noise estimation that appears in the computation of (13), (14) is obtained by a first-order recursion in conjunction with an energy based Voice Activity Detector (VAD) [11].

3. ROBUSTNESS TO SPEAKER VARIATION

In this section, various algorithms for dealing with pronunciation variation in non-native speech are presented. The basic assumption in this work is that non-native speakers tend to confuse phones in the spoken language with phones in their native language. We also assume that information about the phonemic space in the native language exists, e.g., native speech recognition models.

3.1. Confusion Based Acoustic Model Modification

Non-native speakers often produce phones in the spoken language as they would do with similar phones in their native language. By taking into account the acoustic models of the native language, ASR performance can be improved.

Some phones of the spoken language may not have corresponding phones in the native language. For instance, the initial consonant of the English article "the" does not exist in French. Some speakers pronounce this phone like a French "z". Furthermore, there are no diphthongs in French. They may be uttered as a sequence of two French phones, as stated by phonetician experts. Thus, in our new approach, the confusion matrix involves a phone of the spoken language and one or more phones of the native language. The confusion matrix between spoken language phones and sequences of native phones is automatically extracted using the speech recognition systems available for both the spoken and native language.

3.1.1. Confusion extraction

Both spoken language and native language ASR systems are used for confusion extraction. For each utterance of the non-native speech database, we carry out a phonetic alignment using the spoken language ASR system and a phonetic recognition using the native language ASR system. These two time-aligned transcriptions are then compared in order to detect the sequence of native phones that was recognized for each spoken language phone in the utterance. Given a spoken language phone L present in the utterance, the sequence associated with L is composed of native language phones whose time interval has at least 50% overlapping with L's time interval.

The next step is to extract the confusion rules from the above phone and phone sequence associations. Having the count of appearance of each association, the maximum likelihood (ML) estimate of the confusion probability is then computed. Only the confusion rules that have the highest probability are taken into account.

3.1.2. HMM Integration

In this step, the acoustic models of the phones of the spoken language are modified according to the confusion rules extracted from the previous step. For each phone L of the spoken language, a new state path is added to the HMM model of L. These new state paths correspond to the right-hand side of the rules selected according to the previous section. Each phone sequence at the right-hand side of a rule is transformed into a corresponding HMM by concatenating the native phone HMMs in the sequence. Multiple confusion rules for the same spoken language phone are combined in parallel HMM topology.

3.2. Adding Graphemic Constraints

We claim that the pronunciation errors (or variants) a non-native speaker produces depend on graphemes (or the writing of words). The same phone (spoken language) may be mis-pronounced in a different manner depending on the graphemes corresponding to that phone. Thus, the phonetic confusion would be more accurate if graphemic constraints were taken into account. The aim is to automatically extract the graphemes linked to the phones for each word in the dictionary. In [27], graphemic constraints and contexts are used; however, this phone-grapheme alignment is done manually.

3.2.1. Automatic phone-grapheme alignment

Given the writing of a word and its pronunciation, the task here is to find to which graphemes (characters) each phone corresponds. Even though it seems similar, this task is different from building a "grapheme to phone" transducer. In our approach, we use a discrete HMM system to perform this alignment. The CMU dictionary was used to train the HMM system. In this discrete HMM system, the characters (graphemes) are the discrete observations and the phones are the HMMs. The HMMs are single-state discrete HMMs. The trained discrete HMM system can be used to align the graphemes and phones of a word using the Baum-Welch algorithm.

4. EXPERIMENTAL RESULTS

In this section, we present experimental results for the algorithms presented above. The Aurora 2, 3 and/or 4 databases were used for the evaluation; special databases were used for multi-microphone ASR and non-native ASR. In all experiments, hidden Markov models and the HTK toolkit were used for acoustic modeling (with the exception of the speech enhancement experiments where a neural network based recognizer was used).

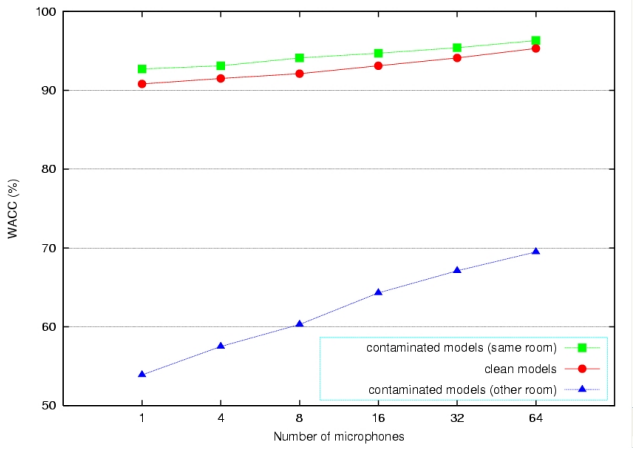


Fig. 1. WACC (%) as a function of array microphones' number.

4.1. Multi-microphone Evaluation

The experimental setup consists of a recognition task of 1001 connected English digits sentences: the original TIDIGITS signals have been reproduced by a loudspeaker in an acoustically treated room (T_{60} is approximately 0.15 s) and acquired by means of a linear array of 64 microphones (Mark III board), located at 1.3m from the simulated speaker. During recordings, in this room, 12 loudspeakers, randomly distributed, were diffusing typical (stationary) cockpit noise. As a result the SNR evaluated at sensors level is about 20dB.

Fig. 1 shows the word accuracy (WACC) of multi-microphone processing as a function of the number of sensor employed in the beamforming using the baseline front-end: 39 MFCCs and cepstral mean subtraction (CMS). and HMM models trained on clean signals (the original TIDIGITS). Two other sets of HMM models have been derived by contamination of the clean training corpus with impulse responses measured in the same room and in another more reverberant room ($T_{60} = 0.7s$), that indeed amplifies the acoustic mismatch. Note that no information regarding the background noise is exploited in this training phase. Overall, up to 50% relative error rate reduction can be achieved by beamforming.

4.2. Robust Features Evaluation

The evaluation experiments are realized by use of the HTK system on Aurora 2 [13] and Aurora 3 (Spanish) databases; (connected-digit recognition, left-right word HMM; Aurora 2: 18-state 3 Gaussian mixtures; Aurora 3: 14-state 16 Gaussian mixtures). The Aurora 2 database contains additive noise in various conditions and SNR. The Aurora 3 database contains recordings from 2 different microphones, at 3 noise conditions that are mixed to create different recognition scenarios - Well-Matched (WM), Medium-Mismatch (MM), High-Mismatch (HM). Apart from the TECCs, the other features are not self-standing, but contribute as 2nd order information to the 1st order speech structure. The input vectors are split in 2 data streams that are assumed independent with stream weights optimized on held-out data. All feature vectors are extended by their time derivatives. The frame length is set equal to 30 ms, with 10ms period update.

As shown in Table 1 by combining MFCCs with AM-FM features we achieve error reduction up to 46% on average on Aurora 3 database. The TECCs outperform the MFCCs especially on the HM scenario on the Aurora 3 task; relative error reduction rate extends up to 73%. The FDCD and MFD methods (Table 2) are evaluated on Aurora 2 showing average relative improvements of 10%, and 15% respectively (improvement at 5 dB SNR MFD: 22%, FDCD: 37%). Compared to the fractal evaluation results, the average performance of the modulation features [6] is similar or slightly better on average on Aurora 2 (average improvement: overall 21%, at 5 dB SNR: 33%). However the modulation features extract different types of nonlinear information than the fractal features.

Scenario / Features	WM	MM	HM	Avg.	Av. Rel. Improv.
Aurora Front-end (WI007)	92.9	80.3	51.6	74.9	-
MFCC+CMS (Baseline)	93.7	92.7	65.2	83.9	36
MFCC+CMS+IA-Mean	93.2	91.4	71.4	85.3	41
MFCC+CMS+IF-Mean	90.7	89.5	72.4	84.2	37
MFCC+CMS+FMP	94.4	92.5	72.8	86.5	46
TECC	93.9	91.8	86.9	90.8	64

Table 1. WACC(%) for Modulation and TECC Features on the Aurora 3 (Spanish Task) Database.

SNR / Feature	clean	20 dB	15 dB	10 dB	5 dB	0 dB
MFCC	98.7	95.7	89.0	71.4	43.5	16.7
+MFD	98.7	96.4	91.6	79.1	52.8	21.7
Improv.	0	1	3	11	22	14
+FDCD	98.58	96.3	92.7	82.9	59.0	22.3
Improv.	0	1	4	16	37	18

Table 2. WACC (%) and Relative Improvement (%) (Improv.) in all tests of the MFCC and the augmented Fractal Features on Aurora 2 (clean training).

4.3. Parametric Equalization Evaluation

Table 3 shows the average word error rate (WERR) of the 14 tests for Aurora 4 clean training experiment. Results are deployed for the BASELINE front end (BASE), for the histogram based equalization (HEQ) , for the proposed parametric equalization (PEQ) and for the ETSI advanced front-end (AFE).

Algorithm	BASE	HEQ	PEQ	AFE
Avg. WERR	45.6	37.5	31.5	31.3

Table 3. WERR (%) for the 14 test Aurora 4 clean training experiment.

4.4. Speech Enhancement Evaluation

Experiments were conducted with a hybrid HMM-NN ASR [10]. The testing conditions used are the following:

- No Denoising (ND): basic Rasta PLP features (RPLP).
- Wiener baseline (WB): RPLP with noise reduction based on standard Wiener filtering.
- Wiener modified (WM): RPLP with Wiener filtering dependent on global SNR [eq. (8)].
- Ephraim-Malah baseline (EMB): RPLP with noise reduction based on the standard Ephraim-Malah spectral attenuation rule [Eqs. (10) (11) (12)].

- Ephraim-Malah modified (EMM): RPLP with noise reduction based on the modified Ephraim-Malah spectral attenuation rule [Eqs. (10) (13) (14)].

The experiment was performed on the Aurora-3 corpus in Italian, Spanish and German, on the High Mismatch test set. The models have been trained with large, domain independent, telephone corpora; the Aurora-3 database was used only for testing. Relevance of results are shown for each test set in parenthesis. Experimental results show that: (i) Ephraim-Malah gain outperforms Wiener gain in its baseline version; this tendency is confirmed when using the modified version of the rules for Wiener gain for Ephraim-Malah gain, and (ii) The modification introduced in the Ephraim-Malah gain produces an average error reduction of 22.9% with respect to the baseline version.

Method	Ita (1.4)	Spa (1.9)	Ger (1.7)	Average
ND	43.3	30.1	17.5	30.3
WB	31.9	18.7	12.2	20.9
WM	25.1	13.8	10.8	16.6
EMB	30.3	18.7	10.3	19.8
EMM	24.4	12.3	9.5	15.4

Table 4. WERR (%) for various speech enhancement algorithms

4.5. Non-Native Speech Evaluation

The non-native database, recorded in the framework of European project HIWIRE, contains 21 French speakers with 100 utterances for each, recorded at a sampling rate of 16 kHz at 16 bits per sample. Each speaker speaks in English. Half of this database was used for development, the other half for testing. The vocabulary is composed of 134 words, and the grammar is a command language. We also used a "word-loop grammar".

system type	WACC	SACC
- baseline system	93.5	87.2
- fully automated "confusion"	**96.1**	**91.1**
- fully automated "confusion" + graphemic confusion	**95.9**	**90.8**
word-loop grammar		
- baseline system	71.1	61.1
- fully automated "confusion"	80.2	66.0
- fully automated "confusion" + graphemic confusion	**81.6**	**67.16**

Table 5. Results for non-native speech (WACC,SACC in %).

Table 5 shows the results of these tests, where "SACC" stands for "sentence accuracy". The "fully automated confusion" (FAC) system achieves a word accuracy of 96.1%, which represents an absolute improvement of 2.6% compared to the "baseline system". The FAC system reduced the WERR by 40% relative. No significant improments were obtained by introducing the graphemic constraints along with the phonetic confusion with constrained grammar. Nevertheless, graphemic constraints allowed further significant improvements when using a word-loop grammar.

5. CONCLUSIONS

A variety of algorithms for improving speech recognition performance in adverse environments and for non-native speakers have been investigated. The experimental results show relative ASR error rate reduction of up to 50% when using beamforming on the signals of multiple microphones. Robust feature extraction methods based on modulation and fractal features improve the recognition accuracy for noisy databases. By combining MFCCs with modulation and fractal features an average error rate reduction of 45% was achieved. It was also shown that the TECC features outperform the MFCCs especially in the presence of additive noise. The proposed parametric feature equalization algorithm is shown to outperform traditional histogram equalization and to reduce error rate by approx. 30%. Speech enchancement is also shown to achieve up to 50% relative error rate reduction and to improve on state-of-the-art speech enhancement algorithms. Finally, the proposed algorithms for dealing with non-native speech show relative error reduction of 30-40%.

Overall, good improvements in ASR performance were achieved under adverse conditions. Future work includes the integration of beamforming and other denoising techniques, adaptive fusion of the different feature streams, extension of the non-native speech models to multiple languages, and evaluation of combinations of the proposed algorithms.

6. REFERENCES

[1] C. Beaugeant, and P. Scalart, "Noise Reduction using Perceptual Spectral Change," in *Proc. Eurospeech*, 1999.

[2] G. Bouselmi, D. Fohr, I. Illina, and J.P. Haton, "Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration," in *Proc. Eurospeech/Interspeech*, 2005.

[3] G. Bouselmi, D. Fohr, I. Illina, and J.P. Haton, "Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration and Graphemic Constraints," in *Proc. ICASSP*, 2006.

[4] M. Brandstein and D. Ward eds., *Microphone Arrays: Techniques and Applications*, Springer, Berlin, 2001.

[5] I. Cohen, " Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, 9,(4):11-117, 2002

[6] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, pp. 621–624, 2005.

[7] ——, "Auditory Teager energy cepstrum coefficients for robust speech recognition," in *Proc. Eurospeech*, 2005.

[8] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum min-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-33, no. 2, pp. 443-445, 1985

[9] L. García, J. C. Segura, J. Ramírez C. Benítez, and A. de la Torre, "Parametric nonlinear feature equalization for robust speech recognition," in *Proc. ICASSP*, 2006.

[10] R. Gemello, D. Albesano, and F. Mana, "Multi-source neural networks for speech recognition," in *Proc. of International Joint Conference on Neural Networks*, Washington, July 1999.

[11] R. Gemello, F. Mana and R. De Mori, "Automatic Speech Recognition with a Modified Ephraim-Malah Rule, *IEEE Signal Proc. Letters*, vol. 13, no. 1, 2006.

[12] J. M. Górriz, J. Ramírez, J. C. Segura, C. G. Puntonet, and L. García, "Effective speech/pause discrimination using an integrated bispectrum likelihood ratio test," in *Proc. ICASSP*, 2006.

[13] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, 2000.

[14] P. Loockwood, and J. Boundy, "Experiments with non-linear Spectral Subtractor (NSS), Hidden Markov Models, and the projection for robust speech recognition in cars," *Speech Communication*, 11, 215-228, 1992.

[15] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: Computation and application to automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1925–1932, 1999.

[16] M. Matassoni, M. Omologo, D. Giuliani,P. Svaizer, "HMM Training with Contaminated Speech Material for Distant-Talking Speech Recognition," *Computer Speech and Language*, 16(2): pp. 205-223. 2002.

[17] J. Morgan, "Making a speech recognizer tolerate non-native speech through Gaussian mixture merging," in *Proc. InSTIL/ICALL*, 2004.

[18] S. Okawa, E. Brocchieri, and A. Potamianos, "Multi–band speech recognition in noisy environments," in *Proc. ICASSP*, 1998.

[19] M. Omologo, P. Svaizer, "Use of the Crosspower-Spectrum Phase in Acoustic Event Location," *IEEE Transactions on Speech and Audio Processing*, 1997.

[20] V. Pitsikalis and P. Maragos, "Speech analysis and feature extraction using chaotic models," in *Proc. ICASSP*, 2002.

[21] V. Pitsikalis and P. Maragos, "Filtered dynamics and fractal dimension for noisy speech recognition," *IEEE Signal Processing Letters*, submitted 2005.

[22] A. Potamianos, E. Sanchez-Soto, and K. Daoudi, "Stream Weight Computation for Multi-Stream Classifiers," in *Proc. ICASSP*, 2006.

[23] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. ICASSP*, 1998.

[24] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[25] J. Ramírez, J. M. Górriz, J. C. Segura, C. G. Puntonet, and A. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum LRT," *IEEE Signal Processing Letters*, 2006.

[26] J. Ramírez, P. Yélamos, J. M. Górriz, and J.C. Segura, "Svm-based speech endpoint detection using contextual speech features," *IEE Electronics Letters*, 2006.

[27] S. Schaden, "Generating non-Native pronunciation lexicons by phonological rule," in *Proc. ICSLP*, 2004.

[28] V. Schless, and F. Class, "SNR-Dependent flooring and noise overestimation for joint application of spectral subtraction and model combination," in *Proc. ICSLP*, 1998.

[29] S. Tamura, K. Iwano, and S. Furui, "A Stream-Weight Optimization Method for Multi-Stream HMMs Based on Likelihood Value Normalization," in *Proc. ICASSP*, 2005.

[30] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez, M. C. Benítez, and A. Rubio, "Histogram equalization of the speech representation for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.