Advanced Front-end for Robust Speech Recognition in Extremely Adverse Environments

Dimitrios Dimitriadis[1], *Jose C. Segura*[2], *Luz Garcia*[2], *Alexandros Potamianos*[3], *Petros Maragos*[1] *and Vassilis Pitsikalis*[1]

[1] School of ECE, National Technical University of Athens, Zografou, Athens 15773, Greece.
[2] Dpto. Teoria de la Senal, Telematica y Comunicaciones (TSTC), Univ. Granada
[3] Dept. of ECE, Technical University of Crete, Chania 73100, Greece
Email:[ddim, maragos, vpitsik]@cs.ntua.gr, [segura, luzgm]@ugr.es, potam@telecom.tuc.gr

Abstract

In this paper, a unified approach to speech enhancement, feature extraction and feature normalization for speech recognition in adverse recording conditions is presented. The proposed front-end system consists of several different, independent, processing modules. Each of the algorithms contained in these modules has been independently applied to the problem of speech recognition in noise, significantly improving the recognition rates. In this work, these algorithms are merged in a single front-end and their combined performance is demonstrated. Specifically, the proposed advanced front-end extracts noise-invariant features via the following modules: Wiener filtering, voice-activity detection, robust feature extraction (nonlinear modulation or fractal features), parameter equalization and frame-dropping. The advanced front-end is applied to extremely adverse environments where most feature extraction schemes fail. We show that by combining speech enhancement, robust feature extraction and feature normalization up to a fivefold error rate reduction can be achieved for certain tasks.

Index Terms: Speech Recognition, Nonlinear Features, Parameter Equalization, Noise Suppression, Noise Invariant Features

1. Introduction

The natural interaction between humans and machines requires the services of robust automatic speech recognition (ASR). The key limitation of current systems is an unreliable level of performance due to the noise conditions in the application environment. For example, in aeronautics, the lack of robustness prevents the more wide-spread introduction of spoken dialogue systems in fixed installations, such as the aircraft cockpit. This application presents strong interest in the aeronautic community as speech-based interaction could provide enhanced safety and efficiency. To create the conditions of an acceptable and robust natural interaction between human and machines, it is necessary to introduce a breakthrough in performance of robust speech understanding. The overall objective of the EU project called *'Human Input That Works In Real Environments'* (HIWIRE) is to set the basis for a much more dependable speech recognition system in the context of noisy environments.

The presence of (several types of) noise in the speech signal can significantly affect the recognition accuracy rates. The performance reduction in the presence of additive noise is due to the contamination of the speech signal and the corresponding change of the feature vectors distributions. A variety of algorithms has been used in the past to improve ASR in noise including signal denoising, speech enhancement, selection of noise-invariant features and feature compensation. Next, we introduce the algorithms that we have selected to include in this advanced front-end, the *Hiwire Advanced Front-End* (HAFE).

The first requirement of our front-end was to feature an efficient noise suppression subsystem consisting of an accurate *Voice Activity Detection* (VAD), working in combination with a *Wiener-based Noise Suppression Filter* (WF) and a *Frame-Dropping* (FD) algorithm. The VAD module could, also, contribute to other key components of robust speech recognition processes like the *Non-linear Feature Normalization*. An identification of the voiced part of speech could significantly contribute to the efficiency of speech detection. With the accurate VAD decisions being possible, frame-dropping is used to remove long non-speech periods from the feature streams. This simple technique has been shown to be very effective, [1].

A Wiener denoising filter is, also, incorporated. Wiener filtering is a signal processing technique that has found a wide applicability under the assumption that the noise is additive, [2]. The estimates for the speech and noise signals require a reliable speech-silence detection process and this introduces the need for an accurate VAD algorithm, too.

The denoised speech signal is the input of the robust feature extraction process. The proposed system yields two different feature streams. The first stream could be either the *Mel Frequency Cepstral Coefficients* (MFCCs) [3], or the *Teager-Energy Cepstral Coefficients* (TECCs) [4], mapping the basic speech structure i.e. the formant structure. On the other hand, the second stream, consisting of either nonlinear modulation or dynamic fractal features, captures the time-varying nature of speech and its micro-structure, [5]. These features provide additional robustness to noise, enhancing further the WF output.

Another popular front-end processing algorithm is feature normalization that attempts to reduce the mismatch between training and operating feature distribution. The problem is to find a transformation that decreases the mismatch between the training (reference) and operating (recognition) environments. The proposed system incorporates a *Parameter Equalization* (PEQ) module where the testing features are equalized according to some statistics computed over the training features.

The paper is organized as follows: First, the system modules are presented in Section 2, producing implementation details and indicating how they interact with the other modules. In Section 3, we describe the speech databases, the speech recognition task setups and finally, the obtained results. The proposed system is evaluated in two different tasks, the Aurora-3 Spanish

Task and the HIWIRE Database task showing significant improvement in ASR performance.

2. Advanced Frontend Modules

The HAFE system consists of the following modules: (i) the raw speech signal processing modules (Wiener filtering, voice activity detection), (ii) the feature extraction module that produces standard MFCCs or TECCs and a variety of nonlinear features motivated by the AM-FM speech model and the theory of fractal speech modeling, (iii) the parametric feature normalization module, (iv) the mean/variance normalization process, and (v) the frame dropping module. The modules operate in cascade, in the order specified. Note that the nonlinear features are used in conjunction with either the MFCCs or the TECCs in a second (independent) feature stream. The various modules are presented next.

2.1. Wiener filter

The first stage of the front-end is a Wiener filter module that performs time domain noise suppression on the speech signal. The filter is designed in the frequency domain as

$$H_t(f) = \left(\sqrt{\xi_t(f)}\right) / \left(1 + \sqrt{\xi_t(f)}\right) \quad (1)$$

where $\xi_t(f) = |\hat{X}_t(f)|^2/|\hat{N}_t(f)|^2$ is an estimate of the *a-priori* SNR, and $|\hat{N}_t(f)|^2$ is an estimate of the background noise power spectrum.

The background noise power spectrum estimate is obtained using a 1^{st}-order recursive filter during non-speech periods

$$|\hat{N}_t(f)| = \lambda|\hat{N}_{t-1}(f)| + (1-\lambda)|Y_t(f)| \quad (2)$$

with $\lambda = 0.99$ and $|Y(f)|$ the magnitude spectrum of the input signal.

The clean speech signal estimate $|\hat{X}_t(f)|$ is obtained in an iterative approach as follows: First, spectral subtraction is used to obtain a first estimation of the clean speech

$$|X_t^1(f)| = \beta|\hat{X}_{t-1}(f)| + (1-\beta)\max(|Y_t(f)| - |\hat{N}_t(f), 0) \quad (3)$$

where $|\hat{X}_{t-1}(f)|$ is the clean signal estimate in the previous frame. Using this estimation and Eq. (1), a first version of the filter $H_t^1(f)$ is obtained; which is used to obtain a better estimate of the clean speech signal $|X_t^2(f)| = H_t^1(f)|Y_t(f)|$. The process is repeated using $|X_t^2(f)|$ and Eq. (1) to get $H_t(f)$ and $|\hat{X}_t(f)| = H_t(f)|Y_t(f)|$. Finally, the frequency domain filter $H_t(f)$ is transformed to the time domain, and applied to the input signal.

2.2. VAD

The voice activity detection module is based on the long-term VAD described in [6]. It is used for the estimation of the background noise characteristics for the Wiener filter design and, also, for the frame-dropping algorithm described in the following sections. The algorithm offers improved speech/non-speech classification accuracy by using contextual temporal information instead of relying on instantaneous power spectrum measures.

The VAD decision is based on the *Long-Term Spectral Divergence* (LTSD) that is defined as

$$LTSD_N(l) = 10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^2(k,l)}{N^2(k)}\right)$$

where $LTSE_N(k,l) = \max\{X(k,l+j)\}_{j=-N}^{j=+N}$ being $X(k,l)$ the amplitude spectrum of the input signal for the k^{th} band of a frame l, and $N(k)$ an estimate of the noise spectrum obtained by averaging over non-speech frames.

2.3. Robust Feature Extraction

The proposed front-end extracts two data streams that are assumed independent. The first one captures the coarse structure of the speech signals, while the second one provides additional information concerning their fine-structure, [5, 7]. The first feature stream consists of either the standard MFCCs or the TECC features (presented in the next section). The second (optional) stream consists of modulation or fractal features presented in Sections 2.3.2 and 2.3.3, respectively.

2.3.1. Teager Energy Cepstral Coefficients (TECC)

The typical MFCCs are estimated over a filterbank of triangular filters with 50% overlap as the log mean squared amplitudes of the bandpass signals, [3]. On the other hand, we propose incorporating information about the time-varying nature of speech using the instantaneous Teager-Kaiser (TK) energy instead of the typical approach. This way, the acoustic information the features' is 'richer'. In addition, we use an auditory-inspired filterbank, [4], instead of the triangular filterbank taking advantage of the human hearing process. The proposed features are shown to be more robust in additive noise and provide additional acoustic information when compared to the MFCCs. These auditory filters are implemented by Gammatone filters and they are smoother and broader than the triangular filters.

The *TECC* estimation algorithm is described with the following steps:

i. Use a Gammatone filterbank to estimate a sequence of bandpass, speech signals. The number of filters is ranging from 25 to 200 filters,

ii. Estimate the mean TK-energy for each one of the framed bandpass signals,

iii. Estimate the Cepstrum coefficients of the log mean energies using DCT, and

iv. Truncate the Cepstrum coefficients to keep the first 13 coefficients (including the 0^{th}-coefficient, c_0).

The first two steps combine the auditory filtering scheme with the more 'natural' approach of the speech TK-energy notion. These steps differentiate the proposed algorithm from the typical MFCC extraction algorithm. The ASR results show significant improvement, especially in noisy recognition tasks, [4].

2.3.2. Modulation Features - FMPs and IFMs

Based on the AM-FM model, [8], a speech signal $s(t)$ can be represented as a sum of a small number N of AM-FM signals $r_i(t)$, where $r_i(t) = a_i(t)\cos\left(\int_0^t f_i(\tau)d\tau\right)$ and $a_i(t)$, $f_i(t)$ (and $i = 1, \ldots, N$) are the Instantaneous Amplitude (IA) and Frequency (IF) modulating signals. Moreover, the nonlinear AM-FM model proposes that the speech resonances $r_i(t)$ are not constant but they can fluctuate around their center frequencies and these fluctuations are mapped onto the IF signals. On the contrary, the linear model of speech assumes that these resonances (and the respective formants' center frequencies) remain constant for relatively short periods of time. This nonlinear model provides additional acoustic information incorporating 2^{nd} -order phenomena that the linear model doesn't capture.

For the decomposition process, we propose using a fixed, Mel-scaled, Gabor filterbank to estimate the bandpassed signals. The filterbank is constant-Q with fixed bandwidth overlap (50%). The Gabor filters are selected due to their optimal span in the Time-Frequency Domains. Finally, we have concluded that the optimal number of filters is 6 when extracting these sets of modulation features, [5].

The *Frequency Modulation Percentages* (FMP) features are defined as $FMP_i = B_i/F_i$ for each speech resonance i, where B_i is the mean bandwidth (an amplitude-weighted version of the $f_i(t)$-signal deviation) and F_i is the weighted mean frequency value of i^{th}-resonance and they provide more accurate and more noise-invariant estimates [9].

In addition to these features, we have examined the use of the *Weighted Mean Inst. Frequency* Coefficients (IFMs), [4]. As mentioned above, formant frequencies are not constant during a single pitch period but they can vary around a center frequency. The IFM coefficients F_i are defined as the (amplitude) weighted mean frequency value of the i^{th}-resonance incorporating some information concerning its fluctuations. The proposed features provide information about the accurate speech formant fine structure, taking advantage of the excellent time-resolution of the ESA, [5]. Transitional phenomena and instantaneous formant variations are mapped onto these FM features. Most often, MFCCs (or even TECCs) fail to capture a significant part the dynamic nature of speech. Thus, we provide this additional information by augmenting the feature vectors with the modulation features, like FMPs or IFMs.

2.3.3. Multiscale Fractal Dimension

The *Multiscale Fractal Dimensions* (MFDs) have been proposed for nonlinear speech analysis and speech recognition in [7, 10]. The main concept is based on the morphological covering algorithm that computes the Minkowski-Boulingand dimension D_M of a planar set. This is computed by dilating the graph of the speech signal with disks B of increasing radii ϵ. If $A_B(\epsilon)$ is the area of the dilated graph, D_M equals $2 - \lim_{\epsilon\to\infty} \log[A_B(\epsilon)]/\log(\epsilon)$. This limit can be estimated from the slope of a line fit to the $\log[A_B(\epsilon)]$ vs $\log(\epsilon)$ data using least squares. The successive local estimates of D_M over moving scale windows yield the MFD.

For the MFD feature set we estimate D_M on the *scalar* speech signals and sample the MFD function at the specific scale values ϵ. We have experimentally observed that the variation of the MFD function is better captured by sampling (at 6 scales) over a logarithmic scale.

2.4. PEQ

Parametric equalization, [11], is a parametric form of the histogram equalization techniques based on a two Gaussian mixture model. The first Gaussian is used to represent non-speech frames, while the second one represents speech frames. For each class, a parametric linear transformation is defined to map the clean and noisy representation spaces,

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

$$\hat{x} = \mu_{s,x} + (y - \mu_{s,y}) \left(\frac{\Sigma_{s,x}}{\Sigma_{s,y}}\right)^{1/2} \quad \text{if } y \text{ is speech} \quad (5)$$

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. The quantities $\mu_{n,y}$, $\Sigma_{n,y}$, $\mu_{s,y}$ and $\Sigma_{s,y}$ correspond to the Gaussians modeling noisy non-speech and speech frames. 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 are estimated from the utterances to be equalized.

To select whether the current frame y is speech or non-speech, the LTSD VAD is used. However, this implies a hard decision between both linear transformations that could create discontinuities in the limit of the non-speech/speech decision. Instead, a soft decision can be used,

$$\hat{x} = P(n|y)\left(\mu_{n,x} + (y - \mu_{n,y})\left(\frac{\Sigma_{n,x}}{\Sigma_{n,y}}\right)^{1/2}\right) + P(s|y)\left(\mu_{s,x} + (y - \mu_{s,y})\left(\frac{\Sigma_{s,x}}{\Sigma_{s,y}}\right)^{1/2}\right) \quad (6)$$

by including the conditional probabilities of frame y being non-speech or speech. The posterior probabilities $P(n|y)$ and $P(s|y)$ are obtained using a simple two-class Gaussian classifier on the log-energy term (the c_0 cepstral coefficient). This classifier is used to obtain the class probabilities $P(n|y)$ and $P(s|y)$ and, also, to obtain the mean and covariance matrices $\mu_{n,y}$, $\Sigma_{n,y}$, $\mu_{s,y}$ and $\Sigma_{s,y}$ for the non-speech and speech classes for the given noisy input utterance. Then, the input utterance can be equalized using Eq. (6). This equation leads to a non-linear interpolation of two class-dependent linear transformations.

2.5. Frame-Dropping - FD

To prevent long non-speech segments to cause insertion errors in the decoding, a simple frame-dropping algorithm is implemented that makes use of the VAD information provided by the LTSD algorithm previously described. Those frames labeled as non-speech by the LTSD VAD are removed from the input stream of the speech recognition engine. To prevent misclassified speech frames to be removed, a simple hang-over algorithm is implemented that delays the VAD decision at the end of speech periods.

3. Experiments

The ASR features are extracted by the proposed front-end system, according to the sequence of preprocessing and post-processing modules mentioned in the previous sections. The nonlinear (modulation or fractal) features are concatenated with either the typical MFCCs or the nonlinear TECCs. The ASR evaluation tasks have been performed on the Aurora-3, Spanish and the HIWIRE Speech databases [12], using the HMM-based HTK Toolkit system, [13]. For the Aurora task, context-independent, 16-state, left-right word HMMs with 3 gaussian mixtures are used. The grammar used is the all-pair, unweighted grammar. In the case of the HIWIRE task, the HMM models are trained on the clean speech TIMIT database and tested on the four different noise-scenario test sets (clean, LN, MN and HN) of the HIWIRE database. The HMM models are 3-state, left-right phone models with 128 mixtures per state. A finite-state grammar with perplexity equal to 14,9 is used as the language model. Finally, the dictionary contains 133 words.

The input vectors are split into two different data streams, one for the standard MFCCs or TECCs and the second one for the nonlinear features. These data streams are assumed statistically independent. The augmented features consist of 13 samples for the 'standard' features (MFCCs/TECCs and their 0^{th}-cepstral coefficient, c_0) and 6 for either the modulation or the fractal features. All feature vectors are extended by their 1^{st} and 2^{nd} time-derivatives and they are smoothed out by *Cepstral Mean Subtraction* (CMS) or *Cepstral Mean and Variance Normalization* (CMVN) to face noise mismatches, additionally to the other denoising techniques. It is shown in [11] that CMS/CMVN schemes in combination with PEQ may improve further the recognition results. Finally, the frame length equals to 30 msec with frame-period equal to 10 msec.

The weights of the two independent data streams are optimized on held-out data. In practice, the stream-weight for the nonlinear features increases with the SNR level, another indication of the robustness of the nonlinear features.

Apart from the baseline features, all the other features have been extracted by the *full HAFE* system using all of its modules (WF, VAD, PEQ, CMS/CMVN and FD), besides the clean-speech case where the WF and PEQ modules are disabled. Wiener filtering smoothes out some part of the acoustic information. So, WF and PEQ modules should be disabled when HAFE is applied to clean speech signals, [2, 11]. However, for all other noise scenarios, the full HAFE should be applied.

Correct Word Accuracy Rates (%) on the Aurora 3, Spanish Task

Features	WM	MM	HM
MFCC (Basel.)	93.68	92.73	65.18
MFCC	96.93	92.98	91.25
TECC	96.90	92.56	91.82
TECC+FMP	97.39	93.75	92.72
TECC+IFM	97.31	94.23	92.81
TECC+MFD	96.98	92.89	92.42
All Features + WF+PEQ+CMS/CMVN+FD			

(a)

Correct Word Accuracy Rates (%) on the HIWIRE Database Task

Features	Clean	LN	MN	HN
MFCC (Basel.)	92.51	45.96	23.31	2.15
MFCC	85.80	69.61	53.82	13.26
TECC	92.80	76.56	53.84	11.81
TECC+FMP	93.86	81.11	61.77	15.61
TECC+IFM	92.13	74.75	58.68	13.45
All Features + WF+PEQ+CMS/CMVN+FD				

(b)

Table 1: Correct Word Accuracy Rates concerning the proposed Frontend. The ASR results are for the (a) Aurora-3, Spanish Task and (b) HIWIRE Task. The baseline features are estimated by the HTK Toolkit.

4. Discussion – Conclusions

In this paper we have presented an advanced front-end system where noise-invariant techniques have been incorporated. This front-end has been applied successfully in extremely adverse environments with significant improvement of the ASR performance. These promising results have been obtained when combining these feature extraction techniques with noise-suppression preprocessing and post-processing modules. We have shown that it is possible to combine heterogeneous subsystems and yield improved recognition results.

5. Acknowledgements

This research work was partially supported by the European Union under the IST-EU STREP program 'HIWIRE'. It was also partially supported by the Greek research program 'Grid-News' of the General Secretariat for Research and Technology.

6. References

[1] J.C. Segura, C. Benitez, A. de la Torre, A. J. Rubio, and J. Ramirez, "Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition," *IEEE Signal Processing Letters*, 2003.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *IEEE Proc.*, 1979.

[3] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, 1980.

[4] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition," in *Eurospeech*, 2005.

[5] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM Features for Speech Recognition," *IEEE Signal Processing Letters*, 2005.

[6] J. Ramirez, J.C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information," *Speech Communication*, 2004.

[7] V. Pitsikalis and P. Maragos, "Filtered Dynamics and Fractal Dimensions for Noisy Speech Recognition," *IEEE Signal Processing Letters*, 2006.

[8] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. on Signal Processing*, 1993.

[9] A. Potamianos and P. Maragos, "Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Demodulation," *J. Acoust. Soc. Am.*, 1996.

[10] P. Maragos and A. Potamianos, "Fractal Dimensions of Speech Sounds: Computation and Application to Automatic Speech Recognition," *J. Acoust. Soc. Am.*, 1999.

[11] L. Garcia, J.C. Segura, J. Ramirez, A. de la Torre, and C. Benitez, "Parametric Nonlinear Feature Equalization for Robust Speech Recognition," in *ICASSP*, 2006.

[12] J.C. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P-A. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos, "The HIWIRE database, A Noisy and Non-native English Speech Corpus for Cockpit Communication," *availabe online at http://www.hiwire.org/*, 2007.

[13] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Ltd., 2002.