Comparative Experiments to Evaluate the Use of Auditory-based Acoustic Distinctive Features and Formant Cues for Robust Automatic Speech Recognition in Low-SNR Car Environments

Hesham Tolba, Sid-Ahmed Selouani & Douglas O'Shaughnessy

INRS-Télécommunications, Université du Québec
800 de la Gauchetière Ouest, Québec, H5A 1K6, Canada
{tolba, selouani, dougo}@inrs-telecom.uquebec.ca

Abstract

This paper presents an evaluation of the use of *some* auditory-based distinctive features and formant cues for robust automatic speech recognition (ASR) in the presence of highly interfering car noise. Comparative experiments have indicated that combining the classical MFCCs with some auditory-based acoustic distinctive cues and either the main formant magnitudes or the formant frequencies of a speech signal using a multi-stream paradigm leads to an improvement in the recognition performance in noisy car environments. To test the use of the new multi-stream feature vector, a series of experiments on speaker-independent continuous-speech recognition have been carried out using a noisy version of the TIMIT database. Using such multi-stream paradigm, we found that the use of the proposed paradigm, outperforms the conventional recognition process based on the use of the MFCCs in interfering noisy car environments for a wide range of SNRs.

1. Introduction

A robust ASR system can be described as a system which can deal with a broad range of applications and adapt to unknown conditions [1]. In general, the performance of existing speech recognition systems, whose designs are predicated on relatively noise-free conditions, degrades rapidly in the presence of a high level of adverse conditions. However, a recognizer can provide good performance even in very noisy background conditions if the exact (same or approximate) testing condition is used to provide the training material from which the reference patterns of the vocabulary are obtained, which is practically not always the case. In order to cope with the mismatched (adverse) conditions, different approaches could be used. The approaches that have been studied for achieving noise robustness can be summarized into two fundamentally different approaches. The first approach attempts to preprocess the corrupted speech input signal prior to the pattern matching in an attempt to enhance the SNR. The second approach attempts to modify the pattern matching itself in order to account for the effects of noise. Methods in this approach include noise masking, the use of robust distance measures, and HMM decomposition. For more details see [2].

In this paper, we focus on optimizing the performance of an ASR system by choosing a suitable distortion measure. The idea of a robust distance measure is to extract relevant features from speech signals which must be insensitive to degradations of the speech signal due to interfering noise or distortions. Many approaches [2] have been used to extract relevant features from a speech signal. Cepstral parameters are well suited to speech recognition due to their compact orthogonality. Unfortunately, cepstral features are highly sensitive to noise. It is well known that cepstral distributions for clean data are well behaved and approximately normal, but in the presence of noise, their profiles are changed significantly and this consequently degrades the performance of an ASR system.

In a previous work, we introduced an auditory-based multi-stream paradigm for ASR [3]. Within this multi-stream paradigm, we merge different sources of information about the speech signal that could be lost when using only the MFCCs to recognize uttered speech. Our experiments showed that the use of some auditory-based features and formant cues via a multi-stream paradigm approach leads to an improvement of the recognition performance. This proves that the MFCCs loose some information relevant to the recognition process despite the popularity of such coefficients in all current ASR systems. In our experiments, we used a 3-stream feature vector. The First stream vector consists of the *classical* MFCCs and their first derivatives, whereas the second stream vector consists of acoustic cues derived from hearing phenomena studies [4]. Finally, the magnitudes of the main resonances of the spectrum of the speech signal were used as the elements of the third stream vector. The above-mentioned work has been extended in [5] by the use of the formant frequencies instead of their magnitudes for ASR within the same multi-stream paradigm. In our experiments, the recognition of speech is performed using a 3-stream feature vector, which uses the formant frequencies of the speech signal obtained through an LPC analysis as the element of the third stream vector combined with the auditory-based acoustic distinctive features and the MFCCs. The obtained results [5] showed that the use of the formant frequencies for ASR in a multi-stream paradigm improves the ASR performance.

In this paper, we extend our work presented in [5] to evaluate the robustness of the above-mentioned proposed features (that is, the acoustic distinctive cues and the formant cues combined with the MFCCs) using a multi-stream paradigm for ASR in noisy car environments. As mentioned above, the MFCCs and their first derivatives, the acoustic cues, which are computed through an auditory-based analysis applied to the speech signal modeled using the Caelen Model [4] and the main four formant frequencies (or magnitudes) of the speech signal obtained through an LPC analysis were combined to form the muti-stream feature vector.

The outline of this paper is as follows. In section 2 we describe the basis of the parameterization process and its importance in ASR. Next, in section 3, we describe briefly the formant extraction procedure. Then in section 4, we proceed with an overview of popular approaches to auditory-based analysis in ASR and we emphasize the description of the auditory Caelen Model that is the basis of the calculation of the proposed acoustic cues for ASR in our study. Next, we describe briefly in section 5 the statistical framework of the multi-stream paradigm. Then in section 6, we proceed with a description of the database, the platform used in our experiments and the evaluation of the proposed approach for ASR. Finally, in section 7 we conclude and discuss our results.

2. Parameterization

The parameterization process in ASR serves to maintain the relevant part of the information within a speech signal while eliminating the irrelevant part for the ASR process. A wide range of possibilities exists for parametrically representing the speech signal such as: short-time spectral envelope, LPC coefficients, MFCCs, short-time energy, zero crossing rates and other related parameters. Among all the parameterization methods, the cepstrum has shown to be favorable for ASR and is widely used in many ASR systems [2].

To better represent temporal variations in the speech signal, higher-order time derivatives (or simply, *delta* parameters for first derivatives, *delta-delta* parameters for the second derivatives) of signal measurements are added to the set of static parameters. The combination of dynamic and static features had shown additional discriminability for speech pattern comparison and consequently improved the accuracy of the speech recognition process. Moreover, temporal variations in the speech signal, obtained by applying time derivatives to the speech signal, when combined with the static features mentioned above, had shown additional discriminability for speech pattern comparison. For more details see [2].

3. Formants

Formant frequencies are defined as the resonance frequencies of the vocal tract. Formants are considered to be representative of the underlying phonetic knowledge of speech and relatively robust in the particular case of ASR in noisy or band-limited environments. However, many problems are associated with the extraction of formants from speech signals. For example, in the case of fricative or nasalized sounds, formants are not well defined. Several methods have been described in the literature provide a solution to the problem of determining formant frequencies. However, accurate determination of formants still poses very difficult problems.

In our experiments, we choose to extract the frequencies of major spectral peak magnitudes using an LPC analysis. This analysis is popular and efficient for representing the spectral envelope with no harmonic effects. A 12-pole-LPC analysis followed by a peak picking algorithm permits us to extract the four frequencies of preeminent peaks, which are considered as *formant-like*. In the context of robust ASR and through our experiments, we investigate if formants provides incremental information to other features used in the proposed multi-stream paradigm.

4. The Auditory-based Processing

4.1. Model-Based Front-End Representation

It was shown through several studies that the use of human hearing properties provides insight into defining a potentially useful front-end speech representation [2]. A filter bank can be regarded as a model of the initial transformation in the human auditory system and often used in speech ASR front-ends. Beside the filter-bank-based techniques, perceptual properties have also been integrated into the analysis of the speech signal through other algorithms such as the *Perceptual Linear Predictive (PLP)* analysis and the so-called *relative spectra* (RASTA) techniques. Including these auditory-based pre-processing techniques in ASR systems led to an improvement of their performances. However, the performance of current ASR systems is far from the performance achieved by humans.

In an attempt to improve the ASR performance in noisy environments, we evaluate in this work the use of the features introduced in [3, 5] for ASR in noisy car environments. These features merge, in a different manner than the methods mentioned above, the hearing/perception knowledge in ASR systems. This is accomplished through the use of the auditory-based acoustic distinctive features and the formant frequencies for robust ASR.

4.2. The Caelen Model

Caelen's auditory model [4] consists of three parts which simulate the behavior of the ear. The external and middle ear are modeled using a bandpass filter that can be adjusted to signal energy to take into account the various adaptive motions of ossicles. The next part of the model simulates the behavior of the basilar membrane (BM), the most important part of the inner ear, that acts substantially as a non-linear filter bank. Due to the variability of its stiffness, different places along the BM are sensitive to sounds with different spectral content. In particular, the BM is stiff and thin at the base, but less rigid and more sensitive to low frequency signals at the apex. Each location along the BM has a characteristic frequency, at which it vibrates maximally for a given input sound. This behavior is simulated in the model by a cascade filter bank. The bigger the number of these filters the more accurate is the model. In front of these stages there is another stage that simulates the effects of the outer and middle ear (pre-emphasis). In our experiments we have considered 24 filters. This number depends on the sampling rate of the signals (16 kHz) and on other parameters of the model such as the overlapping factor of the bands of the filters, or the quality factor of the resonant part of the filters. The final part of the model deals with the electro-mechanical transduction of hair-cells and afferent fibers and the encoding at the level of the synaptic endings. For more details see [4].

4.3. Acoustic Distinctive Cues

The acoustic distinctive cues are calculated starting from the spectral data using linear combinations of the energies taken in various channels. Indeed by such calculations one seeks to describe with these few parameters the spectral distribution and its temporal evolution. It was shown in [6] that 12 acoustic cues are sufficient to characterize acoustically all languages. However, it is not necessary to use all of these cues to characterize a specific language. In our study, we choose 7 cues to be merged in a multi-stream feature vector in an attempt to improve the performance of ASR. These cues are based on the Caelen ear model described above, which does not correspond *exactly* to Jakobson's cues. It was shown in [7] that these acoustic cues are relevant to characterize the different phonemes. Each cue is computed based on the output of the 24 channel filters of the above-mentioned ear model. These seven normalized acoustic cues are: acute/grave (AG), open/closed (OC), diffuse/compact (DC), sharp/flat (SF), mat/strident (MS), continuous/discontinuous (CD) and tense/lax (TL). For more details see [7].

5. Multi-stream Statistical Framework

HMMs constitute the most successful approach developed for modeling the statistical variations of speech in an ASR system. Each individual phone (or word) is represented by an HMM. In large-vocabulary recognition systems, HMMs usually represent subword units, either context-independent or context-dependent, to limit the amount of training data and storage required for modeling words. Most recognizers use typically left-to-right HMMs, which consist of an arbitrary number of states N. The output distribution associated with each state is dependent on one or more statistically independent streams. Assuming an observation sequence $\mathbf{O}$ composed of S input streams $\mathbf{O}_s$ possibly of different lengths, representing the utterance to be recognized, the probability of the composite input vector $\mathbf{O}_t$ at a time t in state j can be written as follows:

$$b_j(\mathbf{O}_t) = \prod_{s=1}^{S} [b_{js}(\mathbf{O}_{st})]^{\gamma_s}, \qquad (1)$$

where $\mathbf{O}_{st}$ is the input observation vector in stream s at time t and γ_s is the stream weight. Each individual stream probability $b_{js}(\mathbf{O}_{st})$ is represented by the most common choice of distribution, *the multivariate mixture Gaussian*:

$$b_{js}(\mathbf{O}_{st}) = \sum_{m=1}^{M} c_{jsm}\, \mathcal{N}(\mathbf{O}_{st}; \mu_{jsm}, \mathbf{\Sigma}_{jsm}), \qquad (2)$$

where M is the number of mixture components in stream s, c_{jsm} is the weight of each mixture component of state j in each mixture of each stream and $\mathcal{N}(\mathbf{O}; \mu, \mathbf{\Sigma})$ denotes a multivariate Gaussian of mean μ and covariance $\mathbf{\Sigma}$ and can be written as:

$$\mathcal{N}(\mathbf{O}; \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} \exp^{-\frac{1}{2}(\mathbf{O}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{O}-\mu)}. \qquad (3)$$

To investigate the multi-stream paradigm using the proposed features for ASR, we have performed a number of experiments in which we merged different sources of information about the speech signal that could be lost with the cepstral analysis. In this paper, we report some experiments in which the parameterization is performed not only using cepstral analysis, but also performed upon both an LPC analysis and an auditory-based analysis applied to the speech signal modeled using the Caelen model described in section 4.2. These experiments have demonstrated a significant performance improvement over the use of the classical MFCCs for ASR in noisy environments. These results are summarized in Table 1.

6. Experiments & Results

6.1. Database & Recognition Platform

In the following experiments the TIMIT database was used. The TIMIT corpus contains broadband recordings of a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States, each reading 10 phonetically rich sentences. To simulate a noisy environment, car noise was added artificially to the clean speech. To study the effect of such noise on the recognition accuracy of the ASR system that we evaluated, the reference templates for all tests were taken from clean speech. On the other hand, the dr1 & dr2 subsets of the TIMIT database were chosen to evaluate the use of the proposed paradigam and features for recognition in noisy car environments.

In order to evaluate our proposed approach for ASR of continuous speech, the HTK-based speech recognition system described in [8] has been used throughout all experiments. The toolkit can be used for isolated or continuous whole-word-based recognition systems. The toolkit was designed to support continuous-density HMMs with any numbers of state and mixture components.

6.2. Tests & Results

In order to evaluate the use of the proposed features for ASR in noisy car environments, we repeated the same experiments performed in our previous studies [3, 5] using the subsets dr1 & dr2 of a noisy version of the TIMIT database at different values of SNR. In these experiments, 12 MFCCs were calculated on a 30-msec Hamming window advanced by 10 msec each frame. Moreover, the normalized log energy is also found, which is added to the 12 MFCCs to form a 13-dimensional (static) vector. This static vector is then expanded to produce a 26-dimensional (static+dynamic) vector. This latter was expanded by adding the seven acoustic distinctive cues that were computed based on the Caelen model analysis. This was followed by the computation of either the main four formant frequencies or their magnitudes, which were added to the MFCCs and the acoustic cues to form a 37-dimensional vector upon which the hidden Markov models (HMMs), that model the speech subword units, were trained. The main four formant frequencies were computed based on an LPC analysis using 12 poles followed by a peak picking algorithm. The proposed system used for the recognition task uses tri-phone Gaussian mixture HMM system.

Five different sets of experiments has been carried out on the noisy version of the TIMIT database. In the first two sets of these experiments, we tested our recognizer using a 30-dimensional feature vector (MFCC_E_D_F & MFCC_E_D_P), in which we combined the magnitudes of the main formant frequencies (or magnitudes) to the classical MFCCs and their first derivatives to form two streams that have been used to perform the recognition process. We found through experiments that the use of these two streams leads to an improvement in the accuracy of the word recognition rate compared to the one obtained when we used the classical 39-dimensional feature vector (MFCC_E_D_A). These tests were performed using N mixture Gaussian HMMs for $N = 1, 2, 4$ & 8 using triphone models for different values of SNR. This is shown in Table 1.

These tests were repeated using the 2-stream feature vector, in which we combined the acoustic distinctive cues to the classical MFCCs and their first derivatives to form two streams (MFCC_E_D_EP) that have been used to perform the recognition process. Again, using these two streams, an improvement in the accuracy of the word recognition rate was obtained for different values of SNR. This is shown in Table 1.

We repeated these tests using the proposed features which combines the MFCCs with the acoustic distinctive cues and the formant frequencies (or magnitudes) to form a three-stream feature

	16 dB	8dB	4 dB	0 dB	-4 dB
MFCC_E_D_A	81.67	58.02	48.02	33.44	22.81
MFCC_E_D_EP	87.60	50.83	38.23	27.29	17.29
MFCC_E_D_P	89.69	69.58	60.73	40.31	27.50
MFCC_E_D_F	88.12	66.25	56.15	38.23	25.31
MFCC_E_D_EP_P	89.38	55.31	41.88	28.44	17.40
MFCC_E_D_EP_F	88.54	51.67	41.04	26.04	20.10

[a] $\%C_{Wrd}$ using 1-mixture triphone models.

	16 dB	8dB	4 dB	0 dB	-4 dB
MFCC_E_D_A	83.85	60.31	49.58	36.56	25.21
MFCC_E_D_EP	88.12	51.98	39.58	28.02	16.56
MFCC_E_D_P	90.21	71.35	59.06	42.92	27.19
MFCC_E_D_F	90.52	67.50	57.50	39.38	27.29
MFCC_E_D_EP_P	89.79	55.73	42.92	29.06	18.12
MFCC_E_D_EP_F	89.58	52.40	41.25	26.25	17.92

[b] $\%C_{Wrd}$ using 2-mixture triphone models.

	16 dB	8dB	4 dB	0 dB	-4 dB
MFCC_E_D_A	84.58	62.40	51.77	35.73	26.25
MFCC_E_D_EP	89.06	53.85	42.29	29.38	17.71
MFCC_E_D_P	89.69	71.67	59.79	42.81	27.81
MFCC_E_D_F	91.35	69.38	60.00	38.75	26.04
MFCC_E_D_EP_P	89.27	58.65	43.75	29.27	19.38
MFCC_E_D_EP_F	90.10	53.54	42.08	26.35	17.40

[c] $\%C_{Wrd}$ using 4-mixture triphone models.

	16 dB	8dB	4 dB	0 dB	-4 dB
MFCC_E_D_A	85.42	63.54	52.60	40.10	28.75
MFCC_E_D_EP	89.38	53.33	41.46	29.27	17.92
MFCC_E_D_P	90.62	70.94	58.85	42.19	28.85
MFCC_E_D_F	90.73	70.00	59.79	40.73	25.00
MFCC_E_D_EP_P	91.35	57.92	43.85	28.75	18.33
MFCC_E_D_EP_F	90.42	54.06	41.04	26.77	19.58

[d] $\%C_{Wrd}$ using 8-mixture triphone models.

Table 1: Comparison of the percent word recognition performance ($\%C_{Wrd}$) of the MFCC_E_D_A-, MFCC_E_D_EP-, MFCC_E_D_P-, MFCC_E_D_F-, MFCC_E_D_EP_P- & MFCC_E_D_EP_F-based HTK ASR systems to the baseline HTK using (a) 1-mixture, (b) 2-mixture, (c) 4-mixture & (d) 8-mixture triphone models and the dr1 & dr2 subsets of the TIMIT database when contaminated by additive car noise for different values of SNR.

vector (MFCC_E_D_EP_P & MFCC_E_D_EP_F). Again, using these combined features, an improvement in the accuracy of the word recognition rate was obtained for different values of SNR. This is shown in Table 1.

7. Conclusion

We have described in this paper an experimental effort to compare the performance of an HMM-based ASR system in noisy car environments when some speech features are combined to the classical MFCCs using a multi-stream paradigm. It was shown through experiments that the use of either the main four formant magnitudes or the frequencies of speech signals, as a side information, combined with some auditory-based features and the classical MFCCs leads to an improvement of the recognition performance of ASR systems in noisy car environments for a wide range of SNR values varying from 16 dB to -4 dB, compared to the systems which use only MFCCs.

Preliminary results show that the inclusion of the above-mentioned features in a multi-stream way (i.e., either the 30 or 33 features which form the 2-stream vector or the 37 features which form the 3-stream vector) reduces the word error rate in noisy car environments for a wide range of SNR values. This shows that the use of the auditory-based acoustic distinctive cues and/or the magnitudes/frequencies of the main spectral peaks (formants) renders the recognition process more robust in noisy car environments. These results shows also that combining a perceptual-based front-end with knowledge gained from measuring the physiological responses to speech stimuli may provide insights into the features used in auditory system for robust speech recognition.

Finally, It should be noted that there were less improvement in the recognition process when we used the formants frequencies than when we used the formants magnitudes for lower values of SNR. This is due to the fact that we did not use a robust algorithm for the formants calculation. We are currently continuing the effort towards the use of robust formant extraction algorithms in an attempt to improve the ASR performance further in very low-SNR environments.

8. References

[1] J-C. Junqua & J-P. Haton, *"Robustness in Automatic Speech Recognition"*, Kluwer Academic Publishers, 1996.

[2] D. O'Shaughnessy, *"Speech Communication: Human and Machine"*, IEEE Press, 2001.

[3] H. Tolba, S.-A. Selouani & D. O'Shaughnessy, *"Auditory-based Acoustic Distinctive Features and Spectral Cues for Automatic Speech Recognition Using a Multi-Stream Paradigm"*, ICASSP'02, May 2002.

[4] J. Caelen, *"Space/Time Data-Information in the ARIAL Project Ear Model"*, Speech Communication, Vol. 4, Nos. 1 & 2, March 1985.

[5] H. Tolba, S.-A. Selouani & D. O'Shaughnessy, *"Comparative Experiments to Evaluate the Use of Auditory-based Acoustic Distinctive Features and Formant Cues for Automatic Speech Recognition Using a Multi-Stream Paradigm"*, ICSLP'02, September 2002.

[6] R. Jakobson, G. Fant & M. Halle, *"Preliminaries to Speech Analysis: The Distinctive Features and their Correlates"*, MIT Press, Cambridge, 1963.

[7] J. Caelen, N. Vigouroux & G. Perennou, *"Structuration des Informations Acoustiques dans le Projet ARIAL"*, Speech Communication, Vol. 2, Nos 2 & 3, July 1983.

[8] Cambridge University Speech Group, "The HTK Book (Version 2.1.1)", Cambridge University Group, March 1997.