Auditory-based Acoustic Distinctive Features and Spectral Cues for Robust Automatic Speech Recognition in Low-SNR Car Environments

Sid-Ahmed Selouani
Université de Moncton
218 bvd. J.-D.-Gauthier,
Shippagan, E8S 1P6, Canada
selouani@umcs.ca

Hesham Tolba
INRS-Télécommunications
800 de la Gauchetière Ouest,
Montréal, H5A 1K6, Canada

Douglas O'Shaughnessy
INRS-Télécommunications
800 de la Gauchetière Ouest,
Montréal, H5A 1K6, Canada

{tolba,dougo}@inrs-telecom.uquebec.ca

Abstract

In this paper, a multi-stream paradigm is proposed to improve the performance of automatic speech recognition (ASR) systems in the presence of highly interfering car noise. It was found that combining the classical MFCCs with some auditory-based acoustic distinctive cues and the main formant frequencies of a speech signal using a multi-stream paradigm leads to an improvement in the recognition performance in noisy car environments.

1 Introduction

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 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 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. For more details see (O'Shaughnessy, 2000).

In a previous work, we introduced an auditory-based multi-stream paradigm for ASR (Tolba et al., 2002). 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. Finally, the magnitudes of the main resonances of the spectrum of the speech signal were used as the elements of the third stream vector.

In this paper, we extend our work presented in (Tolba et al., 2002) to evaluate the robustness of the proposed features (the acoustic distinctive cues and the spectral cues) using a multi- stream paradigm for ASR in noisy car environments. As mentioned above, the first stream consists of the MFCCs and their first derivatives, whereas the second stream vector consists of the acoustic cues are computed from an auditory-based analysis applied to the speech signal modeled using the Caelen Model (Caelen, 1985). Finally, the values of the main peaks of the spectrum of the speech signal were used as the elements of the third stream vector. The magnitudes of the main peaks were obtained through an LPC analysis.

The outline of this paper is as follows. In section 2, an overview on the auditory Caelen Model is given. Next, we describe briefly in section 3 the statistical framework of the multi-stream paradigm. Then in section 4, we proceed with the evaluation of the proposed approach for ASR. Finally, in section 5 we conclude and discuss our results.

2 The Auditory-based Processing

It was shown through several studies that the use of human hearing properties provides insight into defining a potentially useful front-end speech representation (O'Shaughnessy, 2000). 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 hearing/perception knowledge for ASR in noisy car environments. This is accomplished through the use of the auditory-based acoustic distinctive features and the formant frequencies for robust ASR.

2.1 The Caelen's Auditory Model

Caelen's auditory model (Caelen, 1985) 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 (Caelen, 1985).

2.2 Acoustic Distinctive Cues

The acoustic distinctive cues are calculated starting from the spectral data using linear combinations of the energies taken in various channels. It was shown in (Jakobson et al., 1951) 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. 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).

3 Multi-stream Statistical Framework

Most recognizers use typically left-to-right HMMs, which consist of an arbitrary number of states N (O'Shaughnessy, 2000). 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 a *multivariate mixture Gaussian*. 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.

4 Experiments & Results

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. Throughout all experiments the HTK-based speech recognition platform system described in (Cambridge University Speech Group, 1997) has been used. The toolkit was designed to support continuous-density HMMs with any numbers of state and mixture components.

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 study (Tolba et al., 2002) using the subsets dr1 & dr2 of a noisy version of the TIMIT database at different values of SNR which varies from 16 dB to -4 dB. In all our 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 the main spectral peak magnitudes, which were added to the MFCCs and the acoustic cues to form a 37-dimensional vector

	16 dB	8dB	4 dB	0 dB	-4 dB
MFCCEDA	81.67	58.02	48.02	33.44	22.81
MFCCEDE	87.60	50.83	38.23	27.29	17.29
MFCCEDP	89.69	69.58	60.73	40.31	27.50
MFCCEDEP	89.38	55.31	41.88	28.44	17.40

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

	16 dB	8dB	4 dB	0 dB	-4 dB
MFCCEDA	83.85	60.31	49.58	36.56	25.21
MFCCEDE	88.12	51.98	39.58	28.02	16.56
MFCCEDP	90.21	71.35	59.06	42.92	27.19
MFCCEDEP	89.79	55.73	42.92	29.06	18.12

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

	16 dB	8dB	4 dB	0 dB	-4 dB
MFCCEDA	84.58	62.40	51.77	35.73	26.25
MFCCEDE	89.06	53.85	42.29	29.38	17.71
MFCCEDP	89.69	71.67	59.79	42.81	27.81
MFCCEDEP	89.27	58.65	43.75	29.27	19.38

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

	16 dB	8dB	4 dB	0 dB	-4 dB
MFCCEDA	85.42	63.54	52.60	40.10	28.75
MFCCEDE	89.38	53.33	41.46	29.27	17.92
MFCCEDP	90.62	70.94	58.85	42.19	28.85
MFCCEDEP	91.35	57.92	43.85	28.75	18.33

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

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

upon which the hidden Markov models (HMMs), that model the speech subword units, were trained. The main spectral peak magnitudes 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. Three different sets of experiments has been carried out on the noisy version of the TIMIT database. In the first set of these experiments, we tested our recognizer using a 30-dimensional feature vector (MFCCEDP), in which we combined the magnitudes of the main spectral peaks 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 MFCCEDA feature vector, 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 (MFCCEDE). Again, using these two streams, an improvement in the accuracy of the word recognition rate has been obtained when we tested our recognizer using N mixture Gaussian HMMs using triphone models for different values of SNR, Table 1. We repeated these tests using the proposed features which combines the MFCCs with the acoustic distinctive cues and the formant frequencies to form a three-stream feature vector (MFCCEDEP). Again, using these combined features, an improvement in the accuracy of the word recognition rate was obtained, Table 1.

5 Conclusion

We have proposed in this paper a multi-stream paradigm to improve the performance of ASR systems in noisy car environments. Results showed that combining the classical MFCCs with the main formant frequencies of a speech signal using a multi- stream paradigm leads to an improvement in the recognition performance in noisy car environments for a wide range of SNR values varying from 16 dB to -4 dB. These results show that the formant frequencies are relevant for the recognition process not only for clean speech, but also for noisy speech, even at very low SNR values. On the other hand, results showed also that the use of the auditory-based acoustic distinctive cues improves the performance of the recognition process in noisy car environments with respect to the use of only the MFCCs, their first and second derivatives at high SNR values, but not for low SNR values.

References

Hesham Tolba, Sid-Ahmed Selouani and Douglas O'Shaughnessy. 2002. *Auditory-based Acoustic Distinctive Features and Spectral Cues for Automatic Speech Recognition Using a Multi-Stream Paradigm*. IEEE-ICASSP'2002: 837-840.

Jean Caelen. 1985. *Space/Time Data-Information in the ARIAL Project Ear Model*. Speech Communication, 4(1&2): 251-267.

Douglas O'Shaughnessy. 2000. *Speech Communication: Human and Machine*. IEEE Press.

Roman Jakobson, Gunnar Fant and Morris Halle. 1951. *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*. MIT Press, Cambridge.

Cambridge University Speech Group. 1997. *The HTK Book (Version 2.1.1)*. Cambridge University Group.