ROBUST AUTOMATIC CONTINUOUS-SPEECH RECOGNITION BASED ON A VOICED-UNVOICED DECISION

Hesham Tolba *Douglas O'Shaughnessy*

INRS-Télécommunications, Université du Québec
16 Place du Commerce, Verdun (Île-des-Soeurs),
Québec, H3E 1H6, Canada
{tolba, dougo}@inrs-telecom.uquebec.ca

ABSTRACT

In this paper, the implementation of a robust front-end to be used for a large-vocabulary Continuous Speech Recognition (CSR) system based on a Voiced-Unvoiced (V-U) decision has been addressed. Our approach is based on the separation of the speech signal into voiced and unvoiced components. Consequently, speech enhancement can be achieved through processing of the voiced and the unvoiced components separately. Enhancement of the voiced component is performed using an adaptive comb filtering, whereas the unvoiced component is enhanced using the modified spectral subtraction approach. We proved via experiments that the proposed CSR system is robust in additive noisy environments (SNR down to 0 dB).

1. INTRODUCTION

The robustness issue in speech recognition can take on a broad range of problems depending on the environment, channel distortion and stress [8]. In this paper, the degradation of the recognition performance of a large vocabulary automatic speech recognition (ASR) system in low SNR environments is explored. Two types of noises, Gaussian noise and uniform noise, are considered for studying such a degradation. In both cases, noise is added artificially to the speech signal under different SNR levels.

The whole system was designed for noise suppression in contaminated speech. Our design is based on the separation of speech signals into a voiced or an unvoiced part. A V-U decision was incorporated in the front-end of a large vocabulary ASR to classify the speech signal; then these two components were enhanced separately. Our speech enhancement system provides information on the pitch, the spectral envelope and the voicing state of each speech segment. We estimate these parameters and enhance the speech by the modification of such parameters in order to account for the presence of the noise which contaminated the speech signal. To remove the background noise, the voiced component is enhanced using an *Adaptive Comb Filter* (ACF) [9], whereas the unvoiced component is processed using a *Modified Spectral Subtraction* (MSS) [2, 3] approach.

Results indicated that the enhanced speech using this proposed approach is virtually indistinguishable from the original clean speech over a wide range of SNRs, i.e., this approach is robust to additive noise. In addition, such an approach for enhancement provided higher recognition accuracy even at low SNRs.

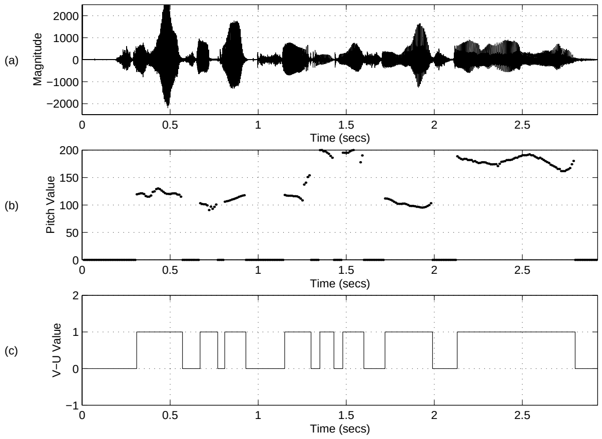

Figure 1: (a) The time-domain waveform of the speech sentence *"She had your dark suit in greasy wash water all year"* uttered by a female speaker from the TIMIT database, (b) Pitch contour (Hz) of the speech sentence shown in (a) estimated using a time-domain algorithm, (c) V-U Classification for the same sentence shown in (a) based on the pitch detection algorithm, all as a function of time.

This paper will be organized into the following sections. Section 2 presents an introduction about the architecture of the proposed speech recognition system and the different parts of the front-end of such a recognizer. In section 3.5, we introduce the feature vector that will be used throughout our experiments in order to test the proposed ASR system. Next, the database and the platform that have been used throughout our experiments are presented in section 3. Following this, experimental results that demonstrate the effectiveness of our proposed approach for recognition are presented in section 4. Finally, in section 5 we conclude and discuss our results.

2. V-U-BASED FRONT-END

The novel proposed ASR system consists of six major parts: voiced-unvoiced classification, enhancement of both voiced and unvoiced components, feature extraction, acoustic/phonetic decoding, lexical access, and syntactic analysis. Operations of these parts will be described in the following subsections.

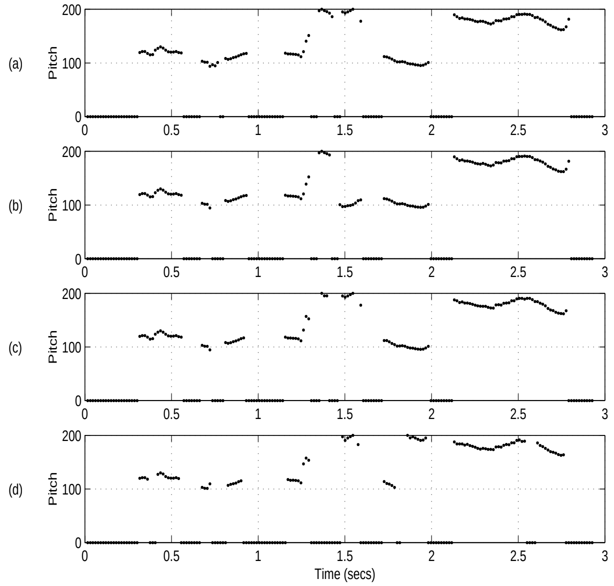

Figure 2: Pitch contour (Hz) of the speech sentence of Fig. 1 for (a) clean speech, (b)-(d) speech contaminated by AGN (average SNR: 27.4716, 16.8301, 11.933 and 6.2125 dB respectively), estimated using a robust time-domain pitch detector.

2.1. Speech Enhancement

Enhancement of the speech signal is achieved by enhancing both the voiced and the unvoiced components of the speech signal separately. The voiced component is enhanced by the use of an *adaptive comb filter* [9], whereas the unvoiced component is processed using the *modified spectral subtraction* approach [2]. The implementation of such an enhancer required three main processing steps prior to the main voiced and unvoiced enhancement processes. These steps are: an accurate estimation of the noise signal, a pitch detector and a V-U classifier. Each of the prior steps will now be discussed.

2.2. Pitch Detection

A pitch detector is necessary for one important reason, which is the use of the ACF in our design. The pitch detector used was based on a time-similarity measure, such as the one introduced in [10]. A typical example of the pitch contour obtained using this approach when applied on a TIMIT database speech file uttered by a female is shown in Fig. 1. Moreover, when this algorithm was applied to this database after adding the Gaussian and the uniform noise, it was found that this algorithm is very robust to such noise and the pitch period calculated is accurate and robust to these noises down to about 0 dB, with a slight change in the estimated pitch.

2.3. V-U Classification

The classification of the speech signal into voiced and unvoiced components provides a preliminary acoustic segmentation of speech, which is important in our design for both speech enhancement and recognition. Different approaches for V-U classifications were described, studied and compared in [12]. Because we deal with noisy speech and the ACF used requires a robust pitch detector to perform the enhancement of the voiced part of the signal, we decided to choose a V-U classifier which is based on the robust pitch detection algorithm described in [10]. After successfully determining the pitch period, a voiced-unvoiced decision was taken, on a frame-by-frame basis, based on a comparison of the correlation values with an adaptive threshold $T(t)$ dependent on the level of the correlation between adjacent pitch periods found for the current segment at that instant [10]. The result of such a classification is shown in Fig. 1. It is clear from this figure that the V-U classification is very accurate even for boundary segments.

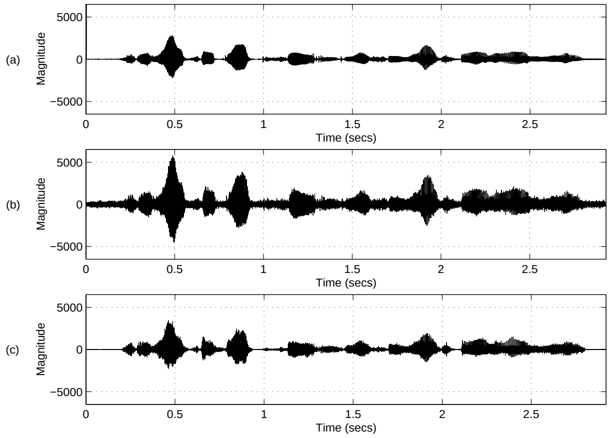

Figure 3: (a) The time-domain waveform of the same speech sentence shown in Fig. 1, (b) Noisy version of the same speech sentence as in (a), with AGN, and the average SNR is 6.21253 dB, (c) Enhanced version of the same speech sentence as in (b) after noise removal using the proposed algorithm ($\alpha = 15$, $\beta = 0.0001$), all as a function of time.

3. EXPERIMENTS

3.1. Database

In the following experiments the TIMIT database [5] was used. To simulate two different types of noise environments, both White Gaussian and uniform noise were added artificially to the clean speech. To study the effect of such noises on the recognition accuracy of the ASR system that we evaluated, the reference templates for all tests were taken from clean speech on the assumption that no *a-priori* noise characteristics knowledge was available. Several separate testing sets were chosen from the available database to evaluate the recognition system. Then, the noise signal was estimated by the detection of the speech pauses to evaluate segments of pure noise. Several methods have been proposed in the literature in order to estimate the noise from the speech corrupted signal [7].

3.2. Noise Estimation and SNR Evaluation

After examining many speech files in the TIMIT database, it was found that the first incoming speech samples of a recording are related to the noise only. Hence, in our experiments, we estimated the noise signal during the first 100 ms of each utterance on a frame-by-frame basis. Then, the average signal energy calculated for such a duration is used as the first estimation of the noise power. After 200 ms, the noise level in a certain subband is estimated by a statistical analysis of a segment of the magnitude spectral envelope. Given a spectral envelope and the corresponding distribution density function in a certain subband, the most frequently occurring spectral magnitude value is taken as an estimation for the noise level inside this band. These noise levels for different subbands are squared and then the average of these squared values gives the noise power estimate. The noise power is computed using this histogram method every 100 ms. More details about such a technique can be found in [7]. Then, the SNR measure, which is based on a frame-by-frame measurement, followed by an averaging over a speech utterance, is used to calculate the SNR per utterance. Moreover, these values are then averaged all over the subset $dr1$ of the TIMIT database to calculate the average SNR for this database.

3.3. Parameter Tuning

A series of experiments at different SNRs, which vary between 30 and 0 dB, have been done in order to determine the optimum value of the parameters of the MSS speech enhancement system, α and β [2], that had been used in the front-end in these experiments. Two types of noise, white Gaussian and Uniform noise, were alternatively added to the clean speech. The values $\alpha = 10$ and $\beta = 0.0001$ were found to be optimal in such experiments using the TIMIT database in order to obtain a more enhanced signal without degrading the naturalness of the speech.

3.4. Speech Reconstruction

In order to test our algorithm, the enhanced speech signal is transformed in the time domain, then it is reconstructed. In general, reconstruction of the speech signal can be performed using the Overlap Add (OLA) procedure described in [1] or the weighted OLA procedure [11, 4].

The speech reconstruction (synthesis) algorithm adopted in our experiments is an OLA algorithm described in [1] using a triangular window. After obtaining the enhanced spectrum of the speech signal, the original noisy phase of the signal spectrum is combined with the enhanced magnitude of the signal, then an IFFT is applied to such a signal to obtain the enhanced signal in the time domain. This time-domain signal is then windowed with a triangular window of length $N_s = 2N$, where N_s is the synthesis-window size and N is the analysis-window size. This windowed segment is then combined with the overlapping portions from neighboring speech segments. Since the distance between successive speech segments is N (30 msec in our experiments), the triangular windows have a 50% overlap (15 msec). The triangular window was found to work well in our experiments; however, this procedure can be generalized to include other window functions.

Fig. 3 illustrates a typical example of a clean speech utterance uttered by a female selected from the subset $dr1$ from the TIMIT database, the same utterance when contaminated by both the additive white Gaussian noise and the enhanced version of such an utterance obtained using the above mentioned enhancement algorithm and reconstructed using the OLA algorithm. This figure indicates that considerable noise rejection has been achieved. The amount of the rejected noise depends on several factors. These factors include: the kind of the noise to be removed; the amount of the SNR level; the optimal choice of the threshold parameters, α and β. Through informal listening testing, it was found that the quality of the enhanced speech is increased; however some residual noise remains.

3.5. Parameterization

The baseline system used for the recognition task was a mono-, bi- and tri-phone Gaussian mixture hidden Markov model (HMM) system. In order to recognize the continuous speech data that has been enhanced as mentioned above, this data is parameterized [6]. 12 Mel frequency cepstral coefficients (MFCCs) are calculated on a 30-msec Hamming window advanced by 10 msec each frame. Then, an FFT is performed to calculate a magnitude spectrum for the frame, which is averaged into 20 triangular bins arranged at equal Mel-frequency intervals. Finally, a cosine transform is applied to such data to calculate the 12 MFCCs. 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 upon which the HMMs, that model the speech subword units, were trained. The static vector is extended by appending the first order difference of the static coefficients. All recognition tests were carried out on the test subset of the TIMIT database. This test set consists of 110 sentences. The data in the TIMIT database was recorded in a clean environment.

4. EVALUATION OF THE V-U-BASED RECOGNIZER IN NOISE

Applying the overall proposed recognizer to the noisy version of the TIMIT database, i.e., after adding both the Gaussian and Uniform noise to the clean signal under different SNRs, which vary between almost 0 and 30 dB, and carrying on some experiments proved that the recognition accuracy has increased significantly, compared to the HTK baseline system.

The results of our evaluation of a subset of the entire database are listed in Tables 2 and 4. These tables show the different recognition error rates for a subset of the TIMIT database when tests were performed using single mixture triphone acoustic models and a word-pair language model using our proposed algorithm. The substitution, deletion and insertion percentage errors were defined respectively as: ϵ_{Sub}, ϵ_{Del} and ϵ_{Ins}, whereas the average word accuracy rate was represented by C_{Ph}.

In order to evaluate the performance of our proposed CSR system, we compared the performance of the V-U-based HTK recognizer to the baseline HTK recognition system. Tables 1 and 3 illustrate the recognition performance obtained using the baseline system when both Gaussian and uniform noises were added to the clean speech for different SNR levels which vary between from almost 4 to 20 dB. It is clear from these results that the V-U-based HTK recognizer outperforms the baseline HTK system and renders the recognition process more robust to additive channel noise. The relative changes in the word correctness rate, C_{Wrd}, when using our proposed system for testing on a subset of the TIMIT database using triphones, the relative changes in C_{Wrd} are 7.23%, 13% and 23.61% when combating AGN for 19.30 dB, 15.91 dB and 11.50 dB SNR levels and 8.39%, 12.22% and 29.81% when combating AUN for 19.58 dB, 15.01 dB and 10.68 dB SNR levels respectively.

SNR	$\epsilon_{Sub}(\%)$	$\epsilon_{Del}(\%)$	$\epsilon_{Ins}(\%)$	$C_{Wrd}(\%)$
19.30 dB	24.09	11.05	0.94	64.86
15.91 dB	27.32	12.51	0.83	60.17
11.50 dB	31.28	18.35	1.15	50.36
7.58 dB	40.77	29.51	0.73	29.72
4.44 dB	43.80	44.32	0.10	11.89

Table 1: Baseline HTK recognition performance versus SNR using single mixture triphones and a subset of the TIMIT database when contaminated by AGN.

SNR	$\epsilon_{Sub}(\%)$	$\epsilon_{Del}(\%)$	$\epsilon_{Ins}(\%)$	$C_{Wrd}(\%)$
19.30 dB	23.04	7.40	2.40	69.55
15.92 dB	25.23	6.78	2.82	67.99
11.50 dB	31.60	6.15	4.59	62.25
7.58 dB	33.68	7.40	5.74	58.92
4.44 dB	47.34	10.32	6.78	42.34

Table 2: V-U-Based HTK recognition performance versus SNR using single mixture triphones and a subset of the TIMIT database when contaminated by AGN.

SNR	$\epsilon_{Sub}(\%)$	$\epsilon_{Del}(\%)$	$\epsilon_{Ins}(\%)$	$C_{Wrd}(\%)$
19.58 dB	25.44	11.26	1.36	63.30
15.01 dB	27.22	13.14	1.04	59.65
10.68 dB	32.74	18.98	1.67	48.28
8.38 dB	39.10	26.69	1.15	34.20
6.90 dB	42.23	31.28	0.52	26.49
5.85 dB	44.11	36.39	0.31	19.50
5.00 dB	43.59	41.40	0.21	15.01

Table 3: Baseline HTK recognition performance versus SNR using single mixture triphones and a subset of the TIMIT database when contaminated by AUN.

SNR	$\epsilon_{Sub}(\%)$	$\epsilon_{Del}(\%)$	$\epsilon_{Ins}(\%)$	$C_{Wrd}(\%)$
19.58 dB	24.09	7.30	3.02	68.61
15.01 dB	26.17	6.88	2.82	66.94
10.68 dB	30.34	6.88	4.69	62.67
8.38 dB	32.33	6.57	6.05	61.11
6.90 dB	38.48	6.88	7.09	54.64
5.85 dB	41.40	8.24	6.36	50.36
5.00 dB	44.73	8.97	6.57	46.30

Table 4: V-U-Based HTK recognition performance versus SNR using single mixture triphones and a subset of the TIMIT database when contaminated by AUN.

5. CONCLUSION

In this paper, a new robust ASR system based on V-U classification has been described. This was realized by the inclusion of such a decision in the pre-processing enhancement algorithm used in the recognition process. We proved via experiments that the proposed CSR system is robust in additive noisy environments and outperforms the baseline recognition system in AGN and AUN environments.

We are currently continuing the effort towards the improvement of the performance of the designed system by modifying the approach that is used for the enhancement of the unvoiced component by the use of an iterative technique such as Wiener filtering.

6. REFERENCES

[1] J. B. Allen, "Short Term Spectral Analysis, Synthesis and Modification by Discrete Fourier Transform", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-25, pp. 235-238, June 1977.

[2] M. Berouti, J. Makhoul and R. Schwartz, "Enhancement of Speech Corrupted by Acoustic Noise", Proc. ICASSP-79, pp. 208-211, 1979.

[3] Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP–27(2), pp. 113–120, April 1979.

[4] R. E. Crochiere, "A Weighted Overlap-Add Method of Short-time Fourier Analysis/Synthesis", IEEE Trans. Acoustics, Speech and Signal Processing", ASSP-28(2), pp. 99–102, February 1980.

[5] William M. Fisher, George R. Dodington and Kathleen M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specification and Status", Proc. DARPA Workshop on Speech Recognition, pp. 93–99, 1986.

[6] Sadaoki Furui, "Speaker–Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP–34(1), pp. 52–59, February 1986.

[7] H. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition", Proc. ICASSP–95, pp. 153–156, 1995.

[8] Jean-Claude Junqua and Jean-Paul Haton, "Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, 1996.

[9] Jae Lim and Alan Oppenheim, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-26(4), pp. 354–358, August, 1978.

[10] Yoav Medan, Eyal Yair and Dan Chazan, "Super Resolution Pitch Determination of Speech Signals", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-39(1), pp. 40–48, January 1991.

[11] M. R. Portnoff, "Time-frequency Representation of Digital Signals and Systems Based on Short-time Fourier Analysis", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-28(2), pp. 55-69, February 1980.

[12] Hesham Tolba, "*Study of Various Inherent Aspects of Robustness and Simplicity of Speech Processing Techniques with Applications to Continuous Speech Recognition in Low-SNR Environments*", Ph. D. Thesis, University of Québec, INRS-Télécommunications, Québec, Canada, 1998.