A time-space adapted wavelet de-noising algorithm for robust automatic speech recognition in low-SNR environments

Hesham Tolba
Electrical Eng. Dept., Faculty of Eng., Alexandria University, Alexandria, Egypt

This paper presents an evaluation of robust large-vocabulary Automatic Speech Recognition (ASR) in the presence of highly interfering car noise using a pre-processing approach based on wavelet-thresholding speech enhancement algorithm that does not require an explicit estimation of the noise level or of the a-priori knowledge of the Speech Noise Ratio (SNR). This algorithm adapts the thresholds in both space and time which allows the removal of various environmental noises. This Time-Space Adapted (TSA) wavelet de-noising algorithm is integrated in the front-end of an ASR system in order to evaluate its robustness in severe interfering car noise environments. The Hidden Markov Model Toolkit (HTK) was used throughout our experiments. Results show that the proposed approach, when included in the front-end of an HTK-based ASR system, outperforms that of the conventional recognition process in severe interfering car noise environments for a wide range of SNRs varying down to 0 dB using a noisy version of the TIMIT database.

هذا البحث يعرض مشكلة كفاءة أنظمة التعرف الأوتوماتيكي علي الكلام عن طريق تحسين إشارة الكلام وذلك قبل بدء عملية التعرف علي الكلام. وذلك باستخدام خوارزميات لا تحتاج إلي معرفة مسبقة بخصائص الضوضاء الموجودة مع الكلام المراد التعرف عليه. هذه الخوارزميات المبنية علي استخدام ال Wavelets تغير من قيم العتبات في كل من الزمان والمكان مما يتيح تنظيف الإشارة من أكبر كم من الضوضاء. هذا الخوارزم تم إضافته إلي مقدمة النظام المستخدم للتعرف الأوتوماتيكي علي الكلام المنطوق في أجواء السيارات العالية الضوضاء. تم استخدام أدوات HTK المعروفة لبناء النظام المقترح للتعرف علي الأصوات، النتائج بينت أن النظام المقترح يفوق الأنظمة العادية وذلك في الأجواء العالية الضوضاء.

Keywords: Speech recognition, Robust speech recognition, Wavelet denoising-algorithms

1. Introduction

The performance of speech recognition systems dramatically decreases when they are trained (in noise-free) and used in different (noisy) environments. 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. One of the major challenges of the speech recognition problem is to make the system robust to background noise [1].

A robust Anosmatic Speech Recogritim (ASR) system can be described as a system which can deal with a broad range of applications and adapt to unknown conditions. 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. Two main approaches to the problem of achieving robust speech recognition in noise can be defined: compensation during the data preprocessing stage, or compensation during the recognition stage. The first approach is classified into two classes. One suppresses the noise component in the speech signal before it is compared with the existing reference patterns in the recognizer. Well-known procedures of this type include Spectral Subtraction or Wiener filtering to remove an estimate of noise from noisy speech observation parameters. The other one is focused on the development of distance measures that are robust to noise contamina-

tion. In this case, there will be no need to create noisy patterns or to process the signal before recognition. The second approach adapts the clean speech models to noise.

In this paper, the first approach of robustness of ASR systems is adopted. The speech enhancement approach that is used to pre-process the speech is based on the TSA wavelet de-noising algorithm that was proposed in [2]. This wavelet thresholding algorithm, which does not require an explicit estimation of the noise level or of the a-priori knowledge of the SNR, adapts the thresholds in both space and time. Such an adaptation allows the removal of various environmental noises and avoids the degradation of speech quality during the thresholding process [2].

This paper is organized as follows. In sections 2 and 3 we describe the basis of the enhancement pre-processing approach that will be integrated in the front-end of our ASR system. Then, we proceed in section 4 with the description of the database, the platform used in our experiments and the evaluation of the recognizer that we are proposing in this paper when used in a noisy car environment, and the comparison of such a recognizer to the baseline recognizer in order to evaluate its performance. Finally, in section 5 we conclude and discuss our results.

2. Denoising by soft thresholding

2.1. Wavelet transform

During the past decade, the Wavelet Transform (WT) has been applied to various research areas. Their applications include signal and image de-noising, compression, detection, and pattern recognition. The WT has recently emerged as a powerful tool for noise reduction [3]. The Wavelet Packet Transform (WPT) [4], which is an extension of the WT, decomposes the signal corrupted with white noise *y(n)* into 2^j subbands corresponding to the wavelet coefficient sets $\omega_{k,m}^j$, where *j* is a given level.

$$\omega_{k,m}^j = WPT\{y(n), j\}, \quad n = 1,2,\ldots,N, \tag{1}$$

$\omega_{k,m}^j$ defines the m^{th} coefficient of the k^{th} subband, where $m = 1,2,\ldots,N/2^j$, $k = 1,2,..2^j$.

2.2. Wavelet shrinkage

The wavelet shrinkage is a simple de-noising technique based on the thresholding of the wavelet coefficients [5]. Assuming that *x(n)* is the noise-free signal and *y(n)* is the signal corrupted with noise *d(n)*, that is:

$$y(n) = x(n) + d(n), \quad n = 1,2,\ldots,N, \tag{2}$$

where *N* is the signal length, we can summarize the de-noising algorithm described by Donoho and Johnston [5,6] as follows:

- WPT of the noisy signal: $\omega_k^j = WPT\{y(n), j\}$.
- Thresholding the resulting wavelet coefficients, to have their shrunken versions $\omega'^j_k = T_s\{\omega_k^j\}$.
- Inverse Wavelet Packet Transform to obtain the enhanced signal, $x'(n) = WPT^{-1}\{\omega'^j_k, j\}$.

In [6], the soft thresholding functions T_s, that have been shown asymptotically near optimal for a wide class of signals corrupted by additive white Gaussian noise, were defined as follows:

$$T_s(\lambda, \omega_k) = \begin{cases} sgn(\omega_k)(|\omega_k| - \lambda) & |\omega_k| > \lambda \\ 0 & |\omega_k| \le \lambda \end{cases}, \tag{3}$$

where ω_k represents the wavelet coefficients and λ is a universal threshold defined as follows:

$$\lambda = \sigma\sqrt{2\log(N)}, \tag{4}$$

where σ = MAD/0.6745 is the noise level and MAD represents the Median Absolute Deviation (MAD) estimated on the first scale. The space-adapted version of this threshold was introduced in [7]. For a given WPT subband *k*, the corresponding threshold is defined by:

$$\lambda_k = \sigma_k\sqrt{2\log(N)}, \quad k = 1,2,\ldots 2^j, \tag{5}$$

where σ_k =MAD$_k$/0.6745 is the noise level and N is the length of the signal. MAD$_k$ represents the MAD estimated on the subband k.

3. Time-Space Adaptation (TSA) using the wavelet transform

In the wavelet shrinkage algorithm proposed by Donoho and Johnston [6,7], the estimated threshold is supposed to define the limit between the wavelet coefficients of the noise and those of the target signal. Unfortunately, it is not always possible to separate the components corresponding to the target signal from those of noise by a simple thresholding. For noisy speech, energies of unvoiced segments are comparable to those of noise. Applying thresholding uniformly to all wavelet coefficients suppresses not only additional noise but also some speech components, such as unvoiced ones. Consequently, the perceptive quality of the filtered speech will be greatly affected.

To prevent speech quality deterioration during the thresholding process, Bahoura and Rouat proposed a new TSA approach for speech enhancement in the wavelet transform domain [2]. Unlike conventional de-noising wavelet methods, the discriminative threshold in various scales is time-adapted as a function of speech components using the Teager Energy Operator (TEO).

3.1. Teager energy approximation

The application of the TEO to the resulting wavelet coefficients $\omega_{k,m}^j$, for a given WPT subband k, led to [2]:

$$t_{k,m}^j = [\omega_{k,m}^j]^2 - \omega_{k,m-1}^j \omega_{k,m+1}^j . \quad (6)$$

This operation enhances the ability to discriminate wavelet coefficients of the speech from those of the noise. Then, an initial mask for each subband k is constructed by smoothing the corresponding TEO coefficients as follows:

$$M_{k,m}^j = t_{k,m}^j * h_k(m), \quad (7)$$

where h_k is a second-order IIR lowpass filter.

3.2. Time-space adapted thresholding

The space-adapted threshold for a given WPT subband k, λ_k, is time-adapted only for speech frames and kept unchanged for non-speech ones. The speech presence is interpreted by significant contrast between peaks and valleys of M_k^j, while its absence is observed with a weak contrast. For each subband k, the time-space adapted threshold is obtained by adapting the corresponding threshold in the time domain according to the following formula:

$$\lambda_{k,m} = \lambda_k (1 - \alpha M'^j_{k,m}), \quad (8)$$

where λ_k is the space-adapted threshold in (5), α is an adjustment parameter, S_k^j is an offset that estimates the valley level to distinguish between speech and non-speech frames and is given by:

$$S_k^j = abscissa[max(H(M_{k,m}^j))], \quad (9)$$

where H is the amplitude distribution of the corresponding mask $M_{k,m}^j$, and $M'^j_{k,m}$ is defined as follows [2]:

$$M'^j_{k,m} = \left[\frac{M_{k,m}^j - S_k^j}{max(M_{k,m}^j - S_k^j)} \right]^{1/8} . \quad (10)$$

The soft thresholding is then applied to the WPT coefficients,

$$\omega'^j_{k,m} = T_s(\lambda_a, \omega_{k,m}^j), \quad (11)$$

where λ_a is the threshold corresponding to the analyzed frame. That is,

$$\lambda_a = \begin{cases} \lambda_{k,m}, & S_k^j \le 0.35\, max(M_{k,m}^j) \\ \lambda_k, & S_k^j > 0.35\, max(M_{k,m}^j). \end{cases} \quad (12)$$

Finally, the enhanced signal $x'(n)$ is synthesized with the inverse WPT of the processed wavelet coefficients as follows [2]:

$$x'(n) = WPT^{-1}\left\{\omega'^{j}_{k,m}, j\right\}. \quad (13)$$

4. Experiments and results

4.1. Database

In the following experiments the TIMIT database was used. The TIMIT corpus of read speech has been designed for the development and evaluation of ASR systems. TIMIT resulted from the joint efforts of several sites under sponsorship from the Defense Advanced Research Projects Agency-Information Science and Technology Office (DARPA-ISTO). Text corpus design was a joint effort among MIT, SRI, and TI. The speech was recorded at TI, transcribed at MIT, and has been verified and prepared for CD-ROM production by the NIST. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States (dr1,dr2,..,dr8). The text material in the TIMIT database consists of 2 dialect sentences designed at SRI, 450 phonetically-compact sentences designed at MIT, and 1890 phonetically-diverse sentences selected at TI. The phonetically-compact sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. Each speaker read 5 of these sentences and each text was spoken by 7 different speakers. The phonetically-diverse sentences were selected to add diversity in sentence types and phonetic contexts. The selection criteria maximized the variety of allophonic contexts found in the texts. Each speaker read 3 of these sentences, with each sentence being read only by a single speaker. The speech material has been divided into portions for training and testing. The test data has a core portion containing 24 speakers, 2 male and 1 female from each dialect region, and 192 sentences. The complete test set contains a total of 168 speakers and 1344 utterances, accounting for about 27% of the total speech material. A Full description of the TIMIT database can be found in [8].

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 proposed, the reference templates for all tests were taken from clean speech.

To evaluate the front-end of the HMM-based recognizer that we propose in this paper, the dr1 subset of the TIMIT database, which consists of about 100 sentences uttered by 10 different speakers (males & females), was chosen from the available database to test the recognition system.

4.2. Recognition platform

In order to recognize the continuous speech data that has been enhanced as mentioned above, the HTK-based speech recognition system described in [9] has been used throughout all the experiments mentioned in this paper. The HTK toolkit can be used for isolated or continuous whole-word-based recognition systems. The HTK toolkit is an integrated suite of software tools for building and manipulating continuous density Hidden Markov Models (HMMs). A HMM can model a specific speech unit such as a subword, a word or a complete sentence. In small-vocabulary recognition systems, HMMs are used to model words. However, in large-vocabulary recognition systems, HMMs usually represent subword units either context-independent (phones) or context-dependent (biphones or triphones), to limit the amount of training data and storage required for modeling words. HMMs constitute the most successful approach developed for modeling the statistical variations of speech. Each individual phone (or word) is represented by a HMM. HTK uses typically left-to-right HMMs, which consist of an arbitrary number of states *N*. The number of states, *N*, can be 5-20 for word models and 5 states for sub-word models in which the entry and the exit states are non-emitting states (i.e., null-states). The output distribution associated with each state is dependent on one or more statistically independent streams.

This toolkit was designed to support continuous-density HMMs with any numbers of state and mixture components. It also implements a general parameter-tying mechanism which allows the creation of complex model topologies to suit a variety of speech recognition applications. It consists of a number of utilities and a comprehensive set of library interface modules. For more details about the HTK toolkit see [9].

The block diagram of the whole recognizer used in the experiments is illustrated in fig. 1 As shown in this figure, the TSA speech enhancing module is also applied to the speech training data before training the HMMs. By training the HMMs of the ASR system with these *modified* speech signals, the HMMs are adapted to the TSA algorithm and therefore some of the distortions due to algorithm can be suppressed. The TSA-adapted version of the clean speech is obtained by simply applying the TSA algorithm to the original clean speech from one of the TIMIT testing sets (dr1) as shown in fig. 1. To construct the noisy speech, the car noise signal is added artificially to the original clean speech at different SNR levels. The enhanced speech is then obtained by enhancing the noisy speech using the TSA algorithm. Finally, the TSA-adapted clean speech and the enhanced speech are tested with the ASR system.

4.3. Test sand results

In all the experiments, 12 MFCCs were 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 39-dimensional vector (including 13 static coefficients, 13 delta coefficients and 13 acceleration coefficients) upon which the hidden Markov models (HMMs), that model the speech subword units, were trained. The baseline system used for the recognition task uses triphone Gaussian mixture HMM system.

Applying the overall proposed recognizer to the noisy version of the TIMIT database under different SNRs, which vary between almost 0 and 20 dB, and carrying on some experiments proved that the recognition accuracy has increased significantly. In order to evaluate the performance of our proposed ASR system, we compared the performance of the wavelet-based HTK recognizer to the baseline HTK recognition system. Table 1 shows a comparison of the percent word correctness rate %C_wrd, recognition accuracy %A_wrd and the degradations in the recognition performance represented by the deletion %E_del, substitution %E_sub and insertion %E_ins percentage errors of the TSA-based HTK ASR system to the baseline HTK using single mixture triphones and the dr1 subset of the TIMIT database when contaminated by additive car noise for different values of SNR. Fig. 2 illustrates the word recognition correctness rates obtained in these ASR tests and table 1 gives some other detailed results. in fig. 2, the dashed line at the top denotes the word recognition correctness rate 95.54% of the clean speech. This can be considered as a baseline compared with that of the noisy speech and the enhanced speech. The second dashed line on the top denotes the word recognition correctness rate 91.98% of the clean speech, when both the training data and testing data have been processed using the TSA algorithm. The lowest dashed line denotes the recognition correctness rates of the noisy speech. It decreases rapidly as the SNR level decreases and shows that ASR performance is sensitive to additive noise. The solid line gives the word recognition correctness rates of the enhanced speech. It is clear from table 1 that the inclusion of the TSA-based wavelet de-noising algorithm in the front-end of our ASR system in noisy car environments reduces the word error rate for a wide range of SNR values down to 0 dB. However, it should be noted that there were no improvement for SNR values greater than 16 dB. This is due to the fact that the parameters that the recognizer uses to recognize the speech signal are altered due to the thresholding process of the wavelet coefficients much greater for high values of

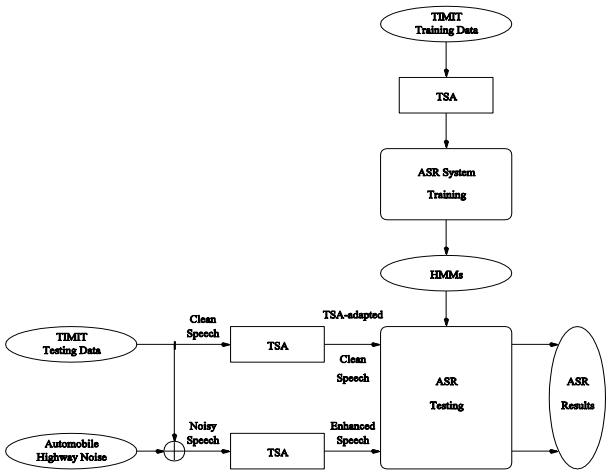

Fig. 1. ASR system for speech enhanced using TSA.

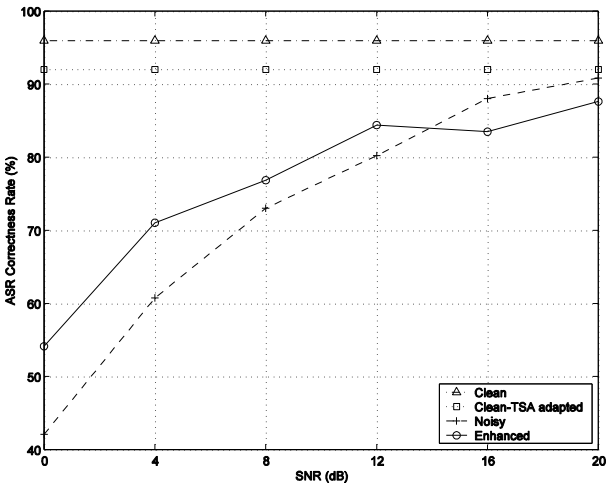

Fig. 2. ASR result s of speech corrupted by car noise and then enhanced using the TSA algorithm.

SNR than for low SNR values. Indeed this limits the use of the TSA-based speech enhancement algorithm to low SNR values.

5. Conclusions

In this paper, the problem of noise robustness of ASR systems using a wavelet-thresholding pre-processing speech enhancement approach was addressed. Preliminary results showed that the inclusion of the TSA

Table 1
Comparison of the percent word recognition performance recognition (%C_wrd), accuracy (%A_wrd), deletion (%E_del), substitution (%E_sub), and insertion (%E_ins) percentage errors of the TSA-based HTK ASR system to the baseline HTK using a noisy version of the TIMIT database when contaminated by additive car noise for different values of SNR

		0dB	4dB	8dB	12dB	16dB	20dB
Noisy	%C_wrd	42.08	60.73	73.02	80.21	88.02	90.83
	%A_wrd	7.71	31.46	51.25	65.42	78.23	85.21
	%E_del	3.54	3.02	1.98	1.25	1.15	0.83
	%E_sub	54.37	36.25	25.0	18.54	10.83	8.33
	%E_ins	34.37	29.27	21.77	14.79	9.79	5.63
TSA-Based	%C_wrd	54.14	71.03	76.85	84.39	83.48	87.61
	%A_wrd	42.36	61.21	68.98	78.67	77.94	83.93
	%E_del	3.96	3.65	1.56	1.88	1.67	0.94
	%E_sub	13.54	11.98	12.08	13.65	10.62	7.08
	%E_ins	5.94	5.31	5.00	5.21	3.65	2.08

wavelet Thresholding algorithm leads to an improvement in the performance of the ASR process in highly interfering car noise environments for a wide range of SNRs down to 0 dB using a noisy version of the TIMIT database. The efforts to improve the performance of the algorithm and to investigate its effects on ASR systems are currently continuing. Although the algorithm is in its preliminary stage, the fact that it requires almost no a-priori knowledge of the noise will certainly lead to an optimistic future.

Acknowledgments

The author thanks Jean Rouat and M. Bahoura for supplying him with the source code of the TSA-based wavelet de-noising algorithm.

References

[1] D. O'Shaughnessy, "Speech Communication: Human and Machine", IEEE Press (2001).

[2] M. Bahoura and J. Rouat, "A New Approach for Wavelet Speech Enhancement", Eurospeech-Scandinavia (2001).

[3] A. Teolis and J. Benedetto, "Noise Suppression Using a Wavelet Model", ICASSP, pp. 17-20 (1994).

[4] R. Coifman, Y. Meyer and M.V. Wickerhauser, "Size Properties of Wavelet Packets", Ruskai et al., pp. 453-470 (1992).

[5] D.L. Donoho, "Nonlinear Wavelet Methods for Recovering Signals, Images, and Densities from Indirect and Noisy Data", Proceedings of Symposia in Applied Mathematics, Vol. 47, pp. 173-205 (1993).

[6] D.L. Donoho, "De-noising by Soft-thresholding", IEEE Trans. Inform. Theory, Vol. 41 (3), pp. 613-627 (1995).

[7] M. Johnston and B.W. Silverman, "Wavelet Threshold Estimators for Data with Correlated Noise", J. Roy. Statist. Soc. B, Vol. 59 (1997).

[8] W. Fisher, G. Doddington and K. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specification and Status", DARPA Workshop on Speech Recognition (1986).

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

Received May 31, 2004
Accepted August 31, 2004