TOWARDS A ROBUST/FAST CONTINUOUS SPEECH RECOGNITION SYSTEM USING A VOICED-UNVOICED DECISION

Douglas O'Shaughnessy *Hesham Tolba*

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

ABSTRACT

In this paper, we show that the concept of Voiced-Unvoiced (V-U) classification of speech sounds can be incorporated not only in speech analysis or speech enhancement processes, but also can be useful for recognition processes. That is, the incorporation of such a classification in a continuous speech recognition (CSR) system not only improves its performance in low SNR environments, but also limits the time and the necessary memory to carry out the process of the recognition. The proposed V-U classification of the speech sounds has two principal functions: (1) it allows the enhancement of the voiced and unvoiced parts of speech separately; (2) it limits the Viterbi search space, and consequently the process of recognition can be carried out in real time without degrading the performance of the system. We prove via experiments that such a system outperforms the baseline HTK when a V-U decision is included in both front- and far-end of the HTK-based recognizer.

1. INTRODUCTION

The performance of CSR systems dramatically decreases when they are trained and used in different environments. However, some features of the speech signal can be extracted accurately even in noisy environments. In this paper, we show that the performance of a CSR system is improved when such features are included in such a system, either in the pre-processing or/and the post-processing stage of such a recognizer. Motivated by this fact, we designed a new system for the automatic recognition of speech signals in highly interfering noise, in which two phases were proposed for the recognition process. The first phase, which is applied at the acoustic level, and is performed by applying a V-U classification of speech signals at the acoustic level, provides us not only with a decision that is used for the enhancement of the speech signal, but also with preliminary information which simplifies the searching complexity during the second phase of recognition at the phonetic-recognition level. This is performed by classifying the speech signal into voiced and unvoiced components using a robust algorithm. These two parts are then enhanced separately using an adaptive comb filter (ACF) and a modified spectral subtraction (MSS) approach to enhance each part. The enhanced signals are then applied to an HMM-based recognizer, the second phase of the recognition, to recognize the spoken utterances via a Viterbi beam search based on the V-U decision information.

While the searching complexity of the proposed system is lowered, the obtained performance of such a system was not as accurate as the original one. This was due to problems that arose from V-U classification errors at the boundaries of the speech segments and the voicing nature of some voiced context-independent phones which are partially devoiced depending on the context. In order to circumvent such problems, some phonetic constraints were added when dealing with the boundary segments. The advantages of our CSR system are the improvement of the recognition performance at low SNRs, as well as the reduction of system complexity. Several modifications to solve the problems associated with this system have also been presented in this paper. These modifications permit our system to work efficiently, even at very low SNRs.

This paper will be organized into the following sections. In section 2, we describe the different parts of the front-end of such a recognizer. Next, in section 3 we proceed with the description of the proposed V-U-based post-processing recognizer, the searching technique and how the V-U decision is included in the post-processing stage. Next, the experimental results that demonstrate the effectiveness of our proposed approach for recognition are presented in section 4. Then, we proceed in section 4 with the evaluation of the proposed front-end and we show how the contaminated signal with severe additive noise is enhanced. Following this, in section 5, the evaluation and the improvement of the V-U-Based recognizer are presented. This is followed, in section 6, by the evaluation and the improvement of the V-U-Based recognizer in noisy environments. Finally, in section 7 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. A simplified block diagram of such a system is illustrated in Fig. 1. These modules were built within the frame of the HTK-based CSR system [12].

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. The V-U decision [10] 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 param-

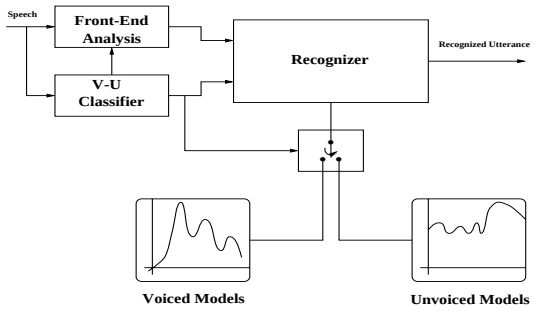


Figure 1: Block diagram of the V-U-based CSR system illustrating how the V-U decision affects both pre-processing and post-processing stages.

eters 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) [8], whereas the unvoiced component is processed using a *Modified Spectral Subtraction* (MSS) [1] approach.

3. V-U-BASED ASR

3.1. V-U Classification

First, consider that speech is classified as voiced or unvoiced on a segment-by-segment basis. Furthermore, consider that the acoustic-modeling is done at the phonetic level, so that we build two classes of probabilistic models for individual phones, α_v and α_u. If the observed segment is classified as voiced, then the probabilistic model class for voiced phones, α_u, will not be considered for recognition and vice-versa.

Our algorithm is a modified version of the well known beam search Viterbi algorithm [11]. In the proposed algorithm, we reduce the searching space using a beam search which is based not only on the score of the different paths at a certain instant, but also on the nature of the nodes, which represent the models of the individual subword units that have been used, at this instant. That is, the chosen nodes at each instant are selected based on the nature of the segments that they present at this instant. To simplify, we divide the models that can model a speech segment as follows:

- $\{\mathbf{V}\overline{\mathbf{U}}\}$ = set of phonemes that can be voiced but cannot be unvoiced.
- $\{\overline{\mathbf{V}}\mathbf{U}\}$ = set of phonemes that can be unvoiced but cannot be voiced.
- $\{\mathbf{VU}\}$ = set of phonemes that can be either voiced or unvoiced.

It must be noted that the $\{\mathbf{VU}\}$ set consists of some of the elements of the $\{\mathbf{V}\overline{\mathbf{U}}\}$ set depending on the context. If a segment of the speech signal is classified as voiced, only the nodes which present the voiced phone models throughout the FSN are selected, and similarly for the unvoiced segments. Consequently, the huge number of nodes of the large vocabulary finite-state network (FSN) is reduced.

3.2. V-U-Based MAP Phonetic Decoder

The task of the phonetic decoder is to extract, from the time series of the acoustic feature vectors, the sequence of phonetic symbols they encode. This operation is performed on the basis of HMMs, where each phonetic unit (phone or allophone) is represented by a first-order HMM. The decoding function is a matter of finding the most likely state sequence of the hidden Markov chain given the observed acoustic feature vectors. This is accomplished using a maximum *a-posteriori* probability (MAP). When the V-U decision is included in our recognizer, the above MAP algorithm is slightly modified by considering *only* some of the paths q_i and not all the paths [10].

The algorithm that we used for computing this sequence is a Listed Viterbi algorithm based on the V-U decision as mentioned above. Consequently, the huge number of models is reduced by dividing the whole system HMM allophonic models into two separate submodels. The benefit is that the number of the models to be trained decreases exponentially and consequently the response time needed for the system also decreases since we do not need to search all models. That is, the incorporation of the V-U decision in the recognition process serves not only for enhancing the corrupted speech and limiting the searching space, but also reduces the number of allophonic models in a large vocabulary recognizer.

4. EXPERIMENTAL RESULTS

4.1. Database and Platform

In the following experiments the TIMIT database [4] was used. The data in the TIMIT database was recorded in a clean environment. To simulate two different types of noise environments, both White Gaussian and uniform noises 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].

The baseline system used for the recognition task was a bi- and tri-phone Gaussian mixture HTK-based [12] system. A modified version of this system, based on the V-U decision, was trained using a 5-state HMM for each phoneme, to define 220 speech states. A single component Gaussian mixture distribution was then trained for each state, for a total of about 34320 parameters. All recognition tests were carried out on the test subset of the TIMIT database. This test set consists of 110 sentences, whereas the training data consisted of 380 sentences from the training set of this database.

4.2. 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 over all the subset $dr1$ of the TIMIT database to calculate the average SNR for this database.

4.3. Parameter Tuning

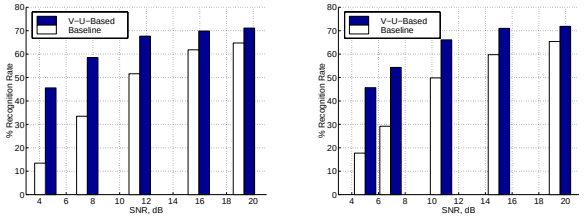
A series of experiments at different SNRs, which vary between 30 and 0 dB, have been done in order to determine the optimum value of the parameters of the MSS speech enhancement system, α and β, that had been used in the front-end in these experiments. Two types of noise, white Gaussian and Uniform noise, 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.

4.4. 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 [10]. 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 [9]. 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 as shown in [9]. The results of such a classification were found to be very accurate even for boundary segments.

4.5. Parameterization

In order to recognize the continuous speech data that has been enhanced as mentioned above, 12 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 as described in [3]. 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 as described in [5].



(a) AGN, 1-mixture triphones

(b) AUN, 1-mixture monophones

Figure 2: Recognition performance comparison as a function of the SNR using single mixture biphones when Gaussian and uniform noises are added to the clean data for different SNR levels.

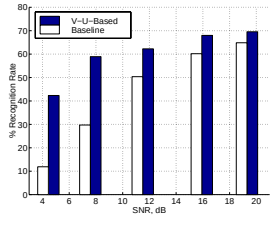
5. EVALUATION AND IMPROVEMENT OF THE V-U-BASED RECOGNIZER

Applying such scheme to the TIMIT database and carrying on some experiments proved that the recognition accuracy dropped compared to the scheme which uses the Viterbi searching algorithm based on high score pruning. This is due to: (1) V-U classification errors and (2) the devoicing of some phonemes dependent on the context. To solve the problem resulting from the misclassification of the speech signal, we proposed to give special care to the boundary segments. This is implemented simply by considering these frames neither voiced nor unvoiced and no pruning is performed, based on the voicing decision for such frames. That is, for such frames all the models, either voiced or unvoiced, were checked and the path that produced the highest score was selected. Doing such a solution, we overcome all the errors that can result from the V-U misclassification errors. The second problem which resulted from the devoicing of some voiced phonemes is solved by adding voicing/devoicing phonetic rules [6]. Some rules depend on the context; however for phones, there is no available information about the dependency of such phonemes on the context. Consequently, two different solutions were proposed to solve the problems for context-independent and -dependent phones, respectively.

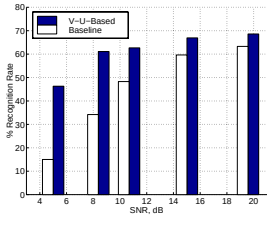
For context-independent phones, the models of the phones which are sometimes devoiced either partially or completely, depending on the context [2] were defined as neutral models, i.e., these models are used for both voiced and unvoiced segments of speech, independent of the classification decision. On the other hand, for context-dependent phones, the voicing/devoicing phonetic rules mentioned above were included in our algorithm and the search is implemented depending on such rules. Combining all the above solutions led to obtaining the same accuracy compared to the original system with a reduction in the search space, while reducing the search complexity.

6. EVALUATION AND IMPROVEMENT 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 Gaussian and uni-



(a) AUN, 1-mixture biphones (RC)



(b) AUN, 1-mixture triphones

Figure 3: Recognition performance comparison as a function of the SNR using single mixture triphones when Gaussian and uniform noises are added to the clean data for different SNR levels.

form noise to the clean signal under different SNRs, which vary between almost 4 and 20 dB, and carrying on some experiments proved that the recognition accuracy has increased significantly as shown in Figures 2 and 3. These figures illustrate the comparison of the recognition performances obtained by the V-U-based and the baseline system using mono-, bi- and tri-phones as speech units respectively, 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 figures 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 single mixture monophones over the baseline HTK, are 18.91%, 28.70% and 62.27% when combating additive Gaussian noise (AGN) for 19.30 dB, 15.91 dB and 11.50 dB SNR levels and 16.83%, 23.19% and 91.38% when combating additive uniform noise (AUN) for 19.58 dB, 15.01 dB and 10.68 dB SNR levels respectively. For right-context (RC) biphones, the relative changes in C_{Wrd} are 9.84%, 12.97% and 31.09% when combating AGN for 19.30 dB, 15.91 dB and 11.50 dB SNR levels and 9.90%, 18.85% and 32.64% when combating AUN for 19.58 dB, 15.01 dB and 10.68 dB SNR levels respectively. For triphones, the relative changes in C_{Wrd} are 7.23%, 13% and 23.61% for the AGN case and 8.39%, 12.22% and 29.81% for the AUN case for the same SNRs as the above tests.

7. CONCLUSION

In this paper, a method of incorporating a V-U decision for a large-vocabulary continuous speech recognizer in noisy environments has been developed, in which we proposed a new front-end analyzer and a new Viterbi beam search decoding architecture that are based on the V-U decision. This decision, which is used as a preliminary method of recognition, adds more constraints to the recognition process. Therefore it was helpful in limiting the search complexity represented in calculation (number of computations) and time. Within the frame work of the HTK, we built such a recognizer which produced, for clean speech, the same performance as the original one which uses the classical Viterbi algorithm, whereas for noisy speech this recognizer outperforms the original one even for low SNRs.

We are currently continuing the effort towards modifying the unvoiced component enhancement approach using an iterative technique such as Wiener filters. Also, the inclusion of a third class (silence) could be very helpful in the enhancement process by attenuating totally the noise in silence regions. This could also reduce the search space further, by considering three categories for the nodes that must be searched throughout the FSN. Finally, the way that we include the V-U decision in our searching algorithm could be changed, i.e., the implementation could be different, which can help in reducing the complexity and getting more accurate performance.

8. REFERENCES

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

[2] Kenneth Ward Church, "Phonological Parsing in Speech Recognition". Kluwer Academic Publishers, 1987.

[3] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-28(4):pp. 357–36, 1980.

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

[5] 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.

[6] Mark Haggard, "The Devoicing of Voiced Fricatives", Journal of Phonetics, 6:pp. 95–102, 1978.

[7] H. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition", Proc. ICASSP–95, pages 153–1566, 1995.

[8] 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.

[9] 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.

[10] 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", PhD thesis, University of Québec, INRS-Télécommunications, Québec, Canada, 1998.

[11] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", IEEE Transactions on Information Theory, IT–13:pp. 260–269, 1967.

[12] Steve Young, P. C. Woodland, and W. J. Byrne, "*HTK: Hidden Markov Model Toolkit V1.5*", Entropic Research Laboratories Inc., Cambridge, 1993.