Robust Romanian Language Automatic Speech Recognizer

DORU-PETRU MUNTEANU, CONSTANTIN-IULIAN VIZITIU
Military Electronic Systems Department
Military Technical Academy
G.Cosbuc 81-83, 050141, Bucharest, Romania
ROMANIA
http://www.mta.ro

Abstract: - In this paper there are presenting solutions for increasing environmental robustness of a Romanian language continuous speech recognizer, previously built [1], [2] as a man-machine dialogue system. Multistyle training strategy is used to train the recognizer with various levels of artificial noise added on the clean speech. Experimental results prove that this scheme strongly increase the system robustness to additive noise.

Key-Words: - continuous speech recognition, environmental robustness, multistyle training

1 Introduction

In this paper it is proposed an architecture for a Romanian language continuous speech recognizer as a human-machine dialogue system and a method for increasing its environmental robustness. The recognizer was built using the multistyle training strategy [3], [4], so that both clean and corrupted speech data were used for training stage.

The majority of speech corpora contain clean speech recorded in low noise reverberation-free conditions. Speech recognition systems performances trained with such clean speech are known to degrade significantly in the real world due to several factors that affect the speech signal such as additive noise (fans, air conditioning, door slams, keyboard or mouse clicks, etc.) or channel distortions (reverberations, microphone frequency response, A/D converter input filter, etc). There are two important strategies for increasing systems robustness: speech enhancement (e.g., spectral noise subtraction, echo cancellation) and acoustical model-based methods (e.g. adaptation techniques, parallel model combination, multistyle training). The speech recognizer proposed in this paper is based on mainly two environmental methods:

- Cepstral mean normalization (CMN) – reduces convolutive channel distortion
- Multistyle training – adapts the models to additive stationary noise

Experimental results prove that system robustness is greatly improved for a wide range of the signal to noise ratio (SNR). Although we have modeled white Gaussian noise only, the method is fitted for modeling any additive noise that could corrupt speech in real-world environments.

2 Speech recognizer architecture

In previously published work [1], [2] it has been described the Romanian language continuous speech recognizer considered. It was build using a very well known toolkit [5] on a Romanian corpus. The main stages of a speech recognizer are presented in Fig. 1. The unknown speech waveform is converted by the acoustic front-end processor into a sequence of acoustic vectors consisting in 12 mel-frequency cepstral coefficients (MFCC) accompanied by their first and second order derivatives.

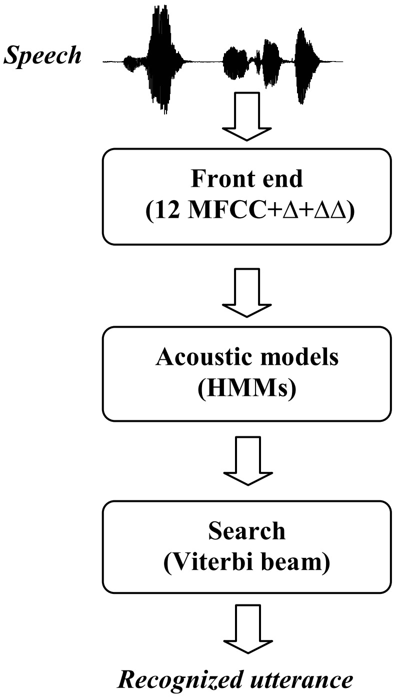

Fig. 1. Continuous Speech Recognition Stages

Phoneme-based context dependent (CD) HMMs with continuous Gaussian distribution were considered. A number of 34 Romanian language phoneme-based context independent (CI) models are trained in the first stage.

It is well known that CI models do not capture the inherent speech variability mainly due to the co articulation effect albeit they are trainable. The recognition system is furthermore refined and first order CD models (triphones) are trained.

3 The acoustical environment

In practice, real world speech differs from clean speech, being degraded by the acoustical environment, which could be defined as the transformations that affects speech from the time it leaves the mouth until it is in digital format. A recognition system is called robust if its accuracy does not degrade too much under mismatched conditions. There are two classes of environmental factors that could corrupt speech:

a) **Additive noise**: computer fans, air conditioning, door slams, other people speech.

b) **Channel distortion**: reverberations, frequency response of the microphone or analog-to-digital converter (CAD).

In most cases, white noise is useful as a conceptual entity, but it seldom occurs in practice. Most of the noise captured by microphones is colored, since its spectrum is not flat (white). For example, pink noise is a particular type of colored noise that has a low-pass nature, as it has more energy at the low frequencies while rolling of at higher frequencies and it could be generated by a computer fan or an automobile engine.

Acoustical environment model is presented in Fig. 2, and the relation between corrupted speech $y[m]$ and clean speech $x[m]$ is given by:

$$y[m] = x[m] * h[m] + n[m] \qquad (1)$$

where $n[m]$ is the additive noise and $h[m]$ is the impulse response of the environment.

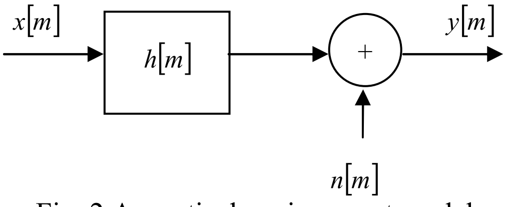

Fig. 2 Acoustical environment model

Regarding the convolution component $h[m]$, the most important factors that could affect the digital form of the speech are reverberation and microphone transfer function. Techniques such as Adaptive Echo Cancellation (AEC) have been successfully applied for reducing the reverberation. The microphone is also very important for the speech acquisition. Head-mounted, close-talking microphones are recommended for most of the speech recognition system as they capture less of the surrounding noise. In order to eliminate the speech variability caused by different digital-analog converters (DAC), it could be included within the head-set and connected by Universal Serial Bus (USB). Another promising strategy for speech acquisition is to use array of microphones. The idea is to use more than one microphone, estimate the relative phase of the signal arriving to each of the element array and than to compute the angle of the arrival. After locating the speaker, all other perturbing signals arriving from other directions or distances are rejected. The major drawbacks of the multi-microphone systems are that they require additional computation to enhance speech and, on the other hand, they also need special hardware (multiple microphones input).

In order to reduce the serious mismatch between the training and test conditions, which often causes dramatic degradation of the accuracy of the recognizers, three major categories of techniques have been developed:

a) Inherently **robust parameters** for speech, such as Perceptual Linear Prediction (PLP)

b) **speech enhancement** including AEC, spectral subtraction (SS), algorithms based on arrays of microphones

c) **model based** methods for noise compensation

In this paper we are presenting experimental results for model based techniques. The major problem for the speech recognizer is the mismatch between the training data (usually, noise-free high quality speech) and test data (environmental conditions).

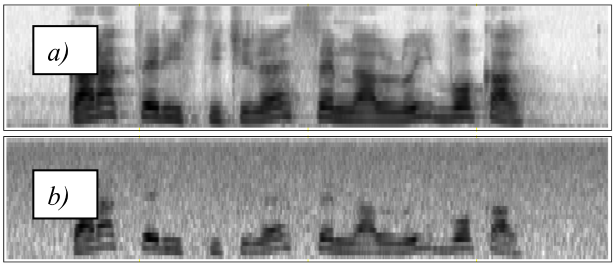

Fig.3 MFCC vs. time for a) clean speech (SNR>40 dB) and b) corrupted speech (SNR=10 dB)

In order to simplify the problem we have referred especially to the additive noise. The simplest approach for this problem is to train the system with the same signal-to-noise ration (SNR) as in the test condition.

The training data may be easily processed by adding to the clean speech noise artificially generated with the same distribution as the noise from the test conditions. Consequently, noisy MFCCs are generated (Fig.3).

Experiments presented here prove that such a matched system performs quite well, much better than the system trained with clean speech, anyway. This simple strategy works *only if the test conditions are known and stationary* but fails in any other situations.

Classical adaptation techniques such as Maximum Posterior Probability Estimation (MAP) or Maximum Likelihood Linear Regression (MLLR) could be used to adapt a clean, speaker-independent recognizer to a particular speaker or to a particular environment. After few thousands of adaptation phrases, the recognition system is adapted to the new condition. Adaptation to new conditions approaches are time consuming and sometimes the applications don't allow them.

In order to increase the environmental robustness of the Romanian language continuous speech recognizer we have adopted the so called multistyle training. Various SNR phrases are produced by adding artificial noise to the clean speech and then the system is trained with the whole collection.

4 Experimental results

4.1 Romanian language continuous speech recognizer

The speech recognizer has an architecture that is described by Fig.1. The acoustical front-end provides 12 mel-frequency cepstral coefficients (MFCC) for each frame of 25 ms, at 100 frames/s rate. Prior to signal parameterization input signal is preemphasized by a filter with the transfer function $H(z) = 1 - 0.97z^{-1}$. Each frame is weighted by a Hamming window. Acoustic vectors are augmented by the first and second variation coefficients.

For acoustical modeling we have used phone-based HMMs with three states in a left-right topology. Continuous Gaussian output distribution with diagonal variance matrices has been adopted. CI models parameters for all 34 Romanian language phonemes were estimated. Then, in order to increase the system accuracy, first-order context-dependent (CD) models, the so-called triphones, have been also trained. We used phonetic decision trees in order to cluster acoustical similar states in a top-down fashion based on data likelihood criteria. Expert knowledge from Romanian language phonetics has been used by means of over 130 phonetic questions in order to determine contextually equivalent classes of HMM states. Training stage was based on uniform model initialization with the global speech mean and variance. Models are than differentiated by the well-known embedded Baum-Welch procedure.

Time-synchronous Viterbi beam search was the strategy for decoding the unknown utterances. Pruning the search space by beam search was very useful for reducing the computation time.

For language modeling (LM), a loop-grammar (Fig. 4.) was adopted, as it is known to be the most difficult task. The reason for choosing this uniform unigram LM is that the system is sensible to any improvements in acoustic modeling.

The system has been trained with a small corpus consisting in 100 phrases uttered by one speaker.

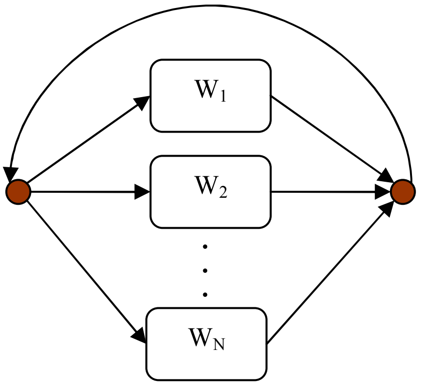

Fig. 4 N-words loop grammar

Recordings were performed with a good quality microphone in noise-free conditions, with a SNR > 40 dB. This clean system has an word error rate (WER) of 14,84 % for monophones and 10,04% for triphones.

4.2 Increasing system robustness

The clean system (trained with clean speech) WER has seriously degraded when we have tested it in mismatch conditions.

For both training and test data we have generated different SNRs phrases in a range between 0 and 25 dB. We have made three groups of experiments for both triphones and monophones:

1. **Clean system**: trained with clean speech tested for each SNR
2. **Matched systems**: trained and tested with the same SNR
3. **Multistyle training**: trained with all phrases (clean + various SNRs) and tested for each SNR

Table 1 WER for the clean system

SNR	WER	
	CI	CD
0	94,65	97,56
5	96,30	95,77
10	83,00	77,00
15	66,86	53,99
20	39,62	29,67
25	22,86	19,15
>40	14,84	10,04

In Table 1, one may see that the clean system performances are quickly degrading as the SNR is decreasing. Obviously, such a system is not robust at all, having a 30 - 40 % WER for normal room conditions with a SNR of 20 dB.

In Fig. 5 and Fig. 6, the results for all three categories of experiments are plotted for CI and CD models, respectively.

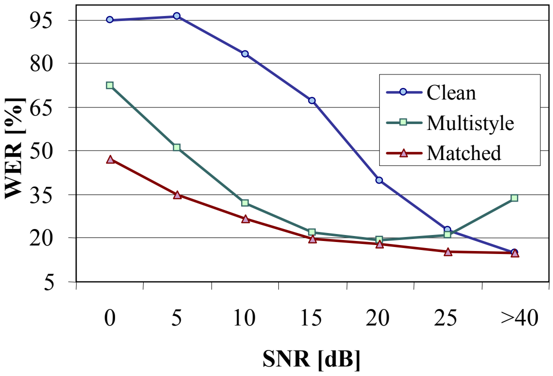

Fig.5 WER in various SNR conditions, for clean system, multistyle training and matched systems (monophones)

Comparative to the matched systems, multistyle training does not require knowledge of the specific noise level and thus is a viable alternative to the theoretical lower bound of matched conditions.

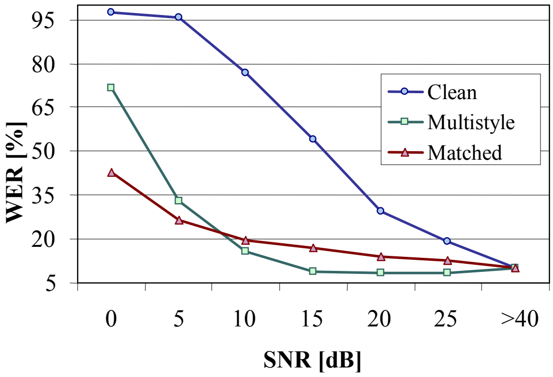

Fig.6 WER in various SNR conditions, for clean system, multistyle training and matched systems (triphones)

One could see the multistyle trained system is clearly more robust than the clean system, being almost as good as the matched system. For the monophone case (Fig.5.), the multistyle system has the best performance in 15-25 dB range as it was trained with 0,5,10,15,20,25 and >40 dB.

The same behavior has the multistyle system for the triphone case (Fig.6.) except that WER is biased below the matched system. Because of the diversity of the training data, the resulting multistyle trained system is more robust to varying noise conditions.

5 Conclusions

In this paper we have presented some solutions for increasing the robustness of a Romanian language continuous speech recognizer previously developed. It is well known that several environmental factors could affect recognition performances in real world man-machine speech based interfaces. In most cases they are critical and the system accuracy is degrading in mismatch conditions. Training the system with phrases at various SNR levels is proved to increase the speech recognition accuracy in a wide range of testing conditions.

References

[1] Munteanu, D., Dumitru, O., Romanian Language Continuous Speech Recognition by Context-Dependency Modeling, *Int. conf. DOGS 2004*, Sombor, Serbia and Montenegro, 2004, pp. 9 – 12.

[2] Oancea, E., Burileanu, C., Munteanu, D., Continuous Speech Recognition System Improvement, *The 3rd Conference on Speech Technology and Human – Computer Dialog (SpeD)*, 2005, pp. 81-91.

[3] Lippmann, R.P., Martin, E.A. and Paul, D.P., Multi-Style Training for Robust Isolated-Word Speech Recognition, *Int. Conf. on Acoustics, Speech and Signal Processing*, Dallas, TX, 1987, pp. 709-712.

[4] Huang, X., Acero A., Hon, H.-W., *Spoken Language Processing – A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

[5] Young, S.J., et all, The HTK Book, Cambridge University Engineering Dept., 2002.

[6] Acero, A., Deng, L., Kristjansson, T. and Zhang, J., Hmm Adaptation Using Vector Taylor Series for Noisy Speech Recognition," Procedings of *ICSLP*, 2000.