Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 49(63), Fascicola 1, 2004

Romanian Language Robust Continuous Speech Recognition

Doru P. Munteanu[1], Eugeniu Oancea[1]

Abstract – In this paper we present some principles for continuous speech recognition, insisting on the influence of several environmental factors that could affect recognition performances. A Romanian language recognizer has been trained using both context dependent and context independent models. Multistyle training strategy was 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.

Keywords: continuous speech recognition, environmental robustness, multistyle training

I. INTRODUCTION

In the last yeas, considerable progress in large vocabulary continuous speech recognition (CSR) has been made. Actual laboratory systems are capable of transcribing continuous speech from any speaker with average error rates under 5%. If speaker adaptation is allowed the error rate could be under 1% after few minutes of speech [9], [10]. Most of these speech recognizers are based on hidden Markov models (HMM) or hybrids HMM-Artificial Neural Networks (ANN). Unfortunately, for practical systems performances are worse because of environmental conditions and the way speakers speak. Robust spontaneous speech recognition is still an elusive goal and actual systems are from far too complex for the performances they are deliver.

This paper discusses principles and architecture of a Romanian language continuous speech recognizer. Experiments performed on a Romanian language task are presented for both clean and noise corrupted speech. The robustness of system was increased by the multistyle training scheme.

II. CONTINUOUS SPEECH RECOGNITION

Speech recognition systems are strongly based on statistical pattern recognition. The main components of a speech recognizer are presented in Fig. 1. An unknown speech waveform is converted by an acoustic front-end processor into a sequence of acoustic vectors. Each of these vectors is a compact representation of the short-time speech spectrum. The recognizer job is to find the best sequence of models for the given sequence of the acoustic vectors. A typical phrase of few seconds could have hundreds of vectors. Search techniques are very important for the system performances. Breadth-first search methods such time-synchronous Viterbi beam search, are more popular than depth-first search like A^*[11]. Language modeling (LM) is another level of the speech recognition process which could greatly improve the recognition rate. For high perplexity tasks, where for instance, words like "pronume" and "prenume" are really difficult to distinguish even for a human listener, information about the context of the word within the phrase is very important.

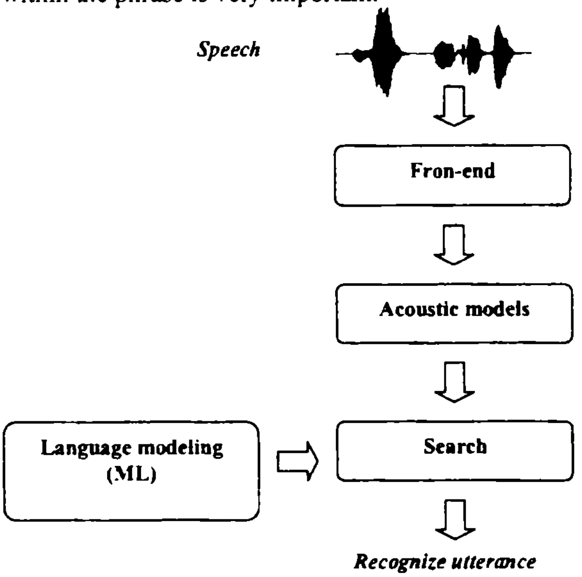

Fig. 1. Continuous Speech Recognizer Architecture

LM based on N-grams is helpful for the recognition process and they could be easily integrated within a decoder. For example, in Viterbi beam search for every transition from the last state of a word w_1 to the first state of the next word w_2, the acoustic score of the current path is increased with the bi-gram probability $p(w_2|w_1)$. Consequently, even if a word

[1] Communication and Electronic Systems Dept., Military Technical Academy, G Coşbuc 81-83, 050141 Bucharest, e-mail: munteanud@mta.ro, eoancea@pcnet.ro

with a high acoustic score and low bi-gram value is to be recognized, the recognizer could choose the word with a bigger bi-gram value and a smaller acoustic score. LM incorporates semantic information into the recognition process and increase the system accuracy. Even with good acoustic and language modeling, practical CSR systems need to be robust to the variability of speech caused by different environmental conditions. Reverberation, additive noise, channel distortions are factors that could seriously degrade the performances of a recognizer that works very well in laboratory.

III. ENVIRONMENTAL ROBUSTNESS

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 its digital form. 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 Figure 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.

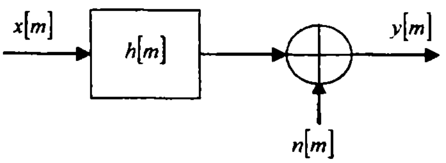

Fig. 2. Acoustical environment model

Regarding the convolutional 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 developed 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 [12]. In order to eliminate the speech variability caused by different digital-analog converters (DAC), this could be included within the head-set and connected by Universal Serial Bus (USB) like in Figure 3. One promising strategy for speech acquisition is to use array of microphones [13], [14]. 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.

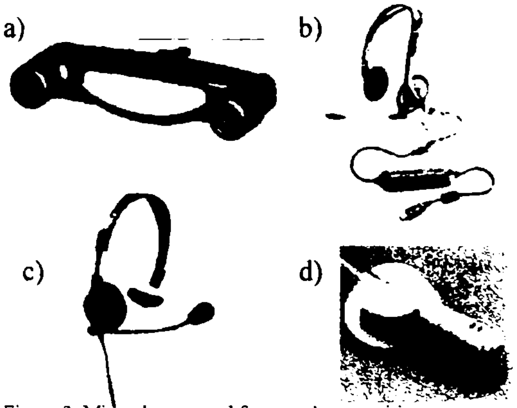

Figure 3. Microphones used for speech recognition: array (a), USB close-talking (b), close-talking(c), desktop(d)

Algorithms for estimating the speaker position based on microphone array could lead to better results in speech recognition even than USB microphones in certain conditions [14]. 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 problem speech recognition designers have to face is the mismatch between the training data (usually, noise-free high quality speech) and test data (environmental conditions). In order to simplify the problem we will refer especially to 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. Our further experiments prove that such a matched system performs quite well, much better than the system trained with clean speech, anyway. This simple strategy works if the test conditions are known and stable but fails in any other situation.

Classical adaptation techniques such as Maximum Posterior Probability Estimation (MAP), Maximum Likelihood Linear Regression (MLLR) or Markov Random Field Linear Regression (MRFLR) could be used to adapt a clean, speaker-independent recognizer to a particular speaker or to a particular environment [8]. After few thousands of adaptation phrases, the recognition system is adapted to the new condition.

Another model-based technique is Parallel Model Combination (PMC) which is based on combining the noise model from the new conditions with the clean models, thus estimating the corrupted speech models. This method needs no additional training or adaptation. After recording few seconds of environmental noise, the clean model is trained and than the new model parameters are computed. Experimental results show that PMC performs quite well in practice [15].

In order to increase the environmental robustness of the Romanian language – continuous speech recognizer (RL-CSR) we have adopted the so called multistyle training. This method is based on producing phrases with various SNR by adding artificial noise to the clean speech and than training the system with the whole collection.

IV. EXPERIMENTAL RESULTS

A. Romanian language – continuous speech recognizer

RL-CSR has an architecture that is described by Figure1. The acoustical front-end provides 12 mel-frequency cepstral coefficients (MFCC) for each frame of 25 ms, at 100 frames/s rate [2]. Before parameterization input signal is pr-emphased 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, left-right topology. Continuous Gaussian distribution with diagonal variance matrices has been adopted. The Romanian language 33 phonemes set was augmented by the post-consonant "i" from the word "pomi". Two models for silence – one for utterance ends with three states and a tee-model with one state for short-pause between words – have been considered. At this point the system was based on context independent models (CI) or monophones.

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 [3], [6], [7]. 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 [4], [5].

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 [1]. Pruning the search space by beam search was very useful for reducing the computation time.

For language modeling (LM), a loop-grammar (Figure 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.

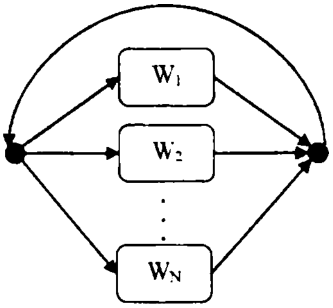

Figure 4. N-words loop grammar

The system has been trained with a small corpus consisting in 100 phrases uttered by one speaker. Recordings were performed with a good quality microphone in noise-free conditions, with a SNR > 30 dB. This clean system has an word error rate (WER) of 14,84 % for monophones and 10,04% for triphones.

B. Increasing system robustness

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

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
50	14,84	10,04

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

In Table 1, one may see that the clean system performances are quickly degrading as the SNR is decreasing. Of course, such a system is impractical, having a 30-40 % WER for normal room conditions with a SNR of 20 dB.

In Figure 5 and Figure 6, the results for all three categories of experiments are plotted for monophones and triphones models. respectively.

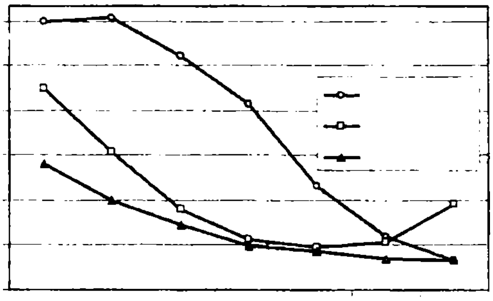

Fig.5. WER in various SNR conditions, for clean system, multistyle training and matched systems (monophones)

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 >30 dB.

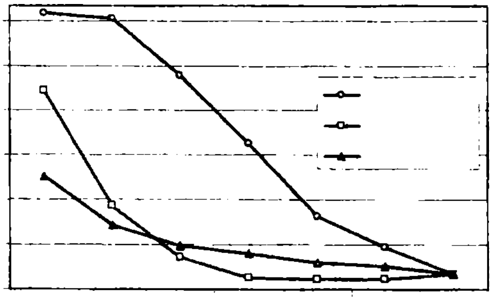

Fig.6. WER in various SNR conditions, for clean system, multistyle training and matched systems (triphones)

The same behavior has the multistyle system for the triphone case (Fig.6.) except that WER is biased below the matched system. The explanation for this bias is that the systems compared have different sizes and the comparison is inaccurate.

V. CONCLUSIONS AND FURTHER WORK

In this paper we have presented some of the CSR principles and architecture. For practical systems, environmental robustness is very important, in most cases critical as the accuracy is degrading in mismatch condition. In order to increase this accuracy, CSR could use inherently robust parameters (PLP), speech enhancement methods (AEC,SS) or model-based techniques (PMC, MAP, MLLR, MRFLR, multistyle training). Experimental results presented herein demonstrate that we can improve the system robustness by simply training him with both clean speec an speec corrupte y no se w t erent SNR. This could work very well in practice if we know the SNR range and the additive noise distribution from the testing conditions.

REFERENCES

[1] D. Munteanu, E. Oancea, "Continuous Speech Recognizer Using Toke- Passing Algorithm", *The 30th Session of Scientific Presentations "Modern Technologies In The XXI Century"*, Bucharest, 2003.

[2] I. Gavat, C.O. Dumitru, G. Costache, "Continuous Speech Recognition Based on Statistical Methods", *Proc SPED 2003*, Bucharest, pp. 115-126.

[3] J.J. Odell, "The Use of Context in Large Vocabulary Speech Recognition", *Dissertation*, University of Cambridge, 1995.

[4] A.I. Rosetti, "Introducere în fonetică", *Ed. Ştiinţifică*, Bucharest, 1967.

[5] E. Oancea, I. Gavăt, O. Dumitru, D. Munteanu, "Improving the Accuracy of the Continuous Speech Recognizers by the Use of Context Dependent Models", *IEEE* Conference „Communications 2004"*, Bucharest, 2004, pp. 221-224.

[6] S.J. Young, "The General Use of Tying in Phoneme-Based HMM Speech Recognizers", *Proc. ICASSP'92*, Vol.1, San Francisco, 1992, pp. 569-572.

[7] .J. Young, J.J. Ode , and P.C. Woodlan , "Tree Base tate Tying for High Accuracy Modeling", *ARPA Workshop on Human Language Technology*, Princeton, 1994.

[8] X.D. Huang, A. Acero, H.-W. Hon, "Spoken Language Processing. A guide to Theory, Algorithm and System Development", *Prentice Hall*, 2001

[9] Placeway, et al., "The 1996 Hub-4 Sphinx-3 System", *Proceedings of the 1997 ARPA Speech Recognition Workshop*, 1997, pp. 85-89.

[10] S.J. Young, "Large Vocabulary Continuous Speech Recognition: A Review", *TR*, Cambridge University, 1996.

[11] D. Mu...a..u, "Alg..i.mi d. căua.. u.ilizaţi în ..cunoaşterea vorbirii", *Referat*, Military Technical Academy, 2003.

[12] D. Munteanu, "Stabilitatea recunoaşterii vorbirii la factori externi", *Referat*, Military Technical Academy, 2004.

[13] J.L. Flanagan, et al., "Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms", *Journal of the Acoustical Society of America*, 78(5), 1985, pp. 1508 - 1518.

[14] M. Schmidt, "Andrea Superbeam® Array Microphone Speech Recognition Performance Using Microsoft Office XP", *Andrea Audio Test Labs White Paper*, 2002.

[15] M.J. Gales, S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination", *IEEE Trans. On Audio and Speech Processing*, 4(5), 1996, pp. 352-359.