Speech Recognition In Non-Stationary Adverse Environments

Zhong-Hua Wang, Patrick Kenny

INRS-Telecommunications, 16 Place du Commerce,
Verdun (Ile-des-Soeurs), Quebec, H3E 1H6, Canada
{wang, kenny}@inrs-telecom.uquebec.ca

Abstract—

In this paper, we introduce a new approach, called nonstationary adaptation (NA), to recognize speech under nonstationary adverse environments. Two models are used: one is a speaker-independent hidden Markov model (HMM) for clean speech, the other is an ergodic Markov chain representing the nonstationary adverse environment. Each state in the Markov chain represents one stationary adverse condition and has associated with it an affine transform that is estimated by maximum likelihood linear regression (MLLR). Three kinds of adverse environments are considered: (i) multi-speaker speech recognition where speaker identity changes randomly and this constitutes a nonstationary adverse condition, (ii) the recognition of speech corrupted by machinegun noise, (iii) the cross-talk problem. The algorithm is tested on the Nov92 development database of WSJF0 with a vocabulary size of 20,000. In multi-speaker speech recognition, NA decreases the error rate by 13.6%. For speech corrupted by machinegun noise, a one-state Markov chain decreases the error rate by 18%, and a two-state Markov chain gives another 14% decrease in error rate. In the cross-talk problem, a one-state Markov chain decreases the error rate by 16.8%. Two-state and three-state Markov chains decrease the error rate by 22% and 24.4%, respectively.

I. Introduction

Current speech recognizers can achieve very good recognition accuracy in the laboratory [1] [2]. However one problem that prevents them from being applied commercially is that they are very sensitive to adverse conditions such as additive noise, channel distortion, speaker variations, etc. [2] [3].

The past two decades have witnessed great advances in the recognition of speech under adverse environments. These approaches can be categorized into robust speech analysis [4], speech enhancement [3][5], the use of robust distance measures [6], model compensation and adaptation [7][8][9][10][11], etc. However most of these techniques concentrate on *stationary* or slowly varying adverse conditions such as car noise, operation room noise, etc.

In this paper, we present a new model combination technique for the recognition of speech under *nonstationary* adverse environments. In this approach, the memory requirement for the corrupted speech model is almost the same as for the clean speech model. Model parameters needed to recognize corrupted speech are calculated on-line at a reasonable computational cost. Two assumptions are made in this approach:

1 A nonstationary adverse environment can be partitioned into a finite number of stationary events and each such event can be represented by one state of a Markov chain.
2 When clean speech is corrupted by a stationary adverse event, the clean speech model parameters can be adapted through an affine transform of the mean vectors.

Based on these two assumptions, we introduce two models: one is a clean speaker-independent HMM that represents the speech signal to be recognized, another one is a Markov chain that represents the nonstationary adverse environment. In the model of the adverse environment, we cannot generally predict which state will follow another one at a specific time; so the corresponding Markov chain is an ergodic one. When clean speech is corrupted by an event that belongs to a state m of the Markov chain, the means of the corrupted speech model are obtained from those of the clean speech model according to the affine transform:

$$\mu(m,c) = \mathbf{A}(m)\mu(c) + \mathbf{B}(m), \qquad (1)$$

where $m = 1, 2, \ldots, M$ and M is the total number of states of the Markov chain, c is the index of a mixture component, $\mu(c)$ is a mean of the clean speech model and $\mu(m,c)$ is a mean of the corrupted speech model corresponding to state m of the noise Markov chain. The variance of the mixture components of the corrupted HMM is assumed to be the same as the clean HMM, since it has been widely reported [9] that variance adaptation gives little improvement in recognition accuracy. We call this approach *nonstationary adaptation* (NA).

NA is applied to three kinds of nonstationary adverse environments. The first is multi-speaker speech recognition where we use a speaker-independent model for clean speech to recognize speech spoken by several known speakers. Since the best recognition result is obtained by using speaker-dependent models, the speaker-independent model introduces some degradation for each speaker compared with a

speaker-dependent model. Such degradation resulting from the ignorance of the speaker identity constitutes the adverse condition. Since speaker identity changes randomly, the adverse condition is nonstationary. The second is the recognition of speech corrupted by machinegun noise [13], which is a typical nonstationary noise. The third is the cross-talk problem where the background noise is a speech signal, which is a quite nonstationary signal.

II. Implementation of the algorithm

A very large vocabulary continuous speech recognition system usually includes two passes. In the first pass, a coarse language model and a set of coarse acoustic phonetic models are used to generate all recognition hypotheses that have high scores. These recognition hypotheses are represented by a *word graph*. There is a one-to-one correspondence between paths through the word graph and recognition hypotheses. In the second pass, a fine language model and a set of fine acoustic phonetic models are used to rescore all these hypotheses. The search space of the second pass is the word graph that was generated in the first pass.

The experiments described in this paper are second pass experiments, using a fixed word graph for each test sentence. The acoustic phonetic model in the second pass is adapted to the corrupted speech model via the set of affine transformations corresponding to the Markov chain of the adverse environment.

Let l, m denote states in the Markov chain, and let i, j denote states in the clean speech HMM. The forward probability, $\alpha(i,m,t)$, of the observation sequence $\mathbf{o}_1 \cdots \mathbf{o}_t$, the HMM state i, and the Markov chain state m is calculated by

$$\alpha(j,m,t) = \sum_i \sum_l \alpha(i,l,t-1) \cdot a_{ij} \cdot c_{lm} \cdot b_{jm}(\mathbf{o}_t)$$

where a_{ij} and c_{lm} are state transition probabilities of the speech model and the Markov chain, respectively, and $b_{jm}(\mathbf{o}_t)$ is the probability of emitting vector $\mathbf{o}_t$ when the clean speech state j is corrupted by an event that belongs to the Markov chain state m. The calculation of $b_{jm}(\mathbf{o}_t)$ is the same as in clean speech recognition except that the mean vector $\mu(c)$ is replaced by $\mu(m,c)$, as illustrated by equation (1). In Viterbi decoding, the forward probability is calculated according to

$$\alpha(j,m,t) = \max_i \max_l [\alpha(i,l,t-1) \cdot a_{ij} \cdot c_{lm} \cdot b_{jm}(\mathbf{o}_t)].$$

In the Markov chain, we assume all intra-state transition probabilities to be one and all inter-state transition probabilities to be a constant C whose value can be either larger or smaller than 1. If C is less than 1, it encourages intra-state transitions and discourages inter-state transitions. We call $\log C$ to be an *inter-state transition penalty* or a briefly just *penalty*.

A. Computational overhead

In the INRS continuous speech recognition system, all distribution components share a full covariance matrix. In noisy speech recognition, the rotation matrix of the affine transforms can be absorbed by the covariance matrix and the likelihood expression is almost the same as in the clean speech recognition. If the adverse environment is represented by an M-state Markov chain, each time a frame is read, a score vector of M components has to be calculated with each component representing a score if the speech is corrupted by an event belonging to one state of the Markov chain. So *likelihood computation* overhead is M times as much as in clean speech recognition. While maintaining reasonable computation overhead, NA keeps the memory requirement almost the same as for clean speech recognition.

B. Adaptation parameter estimation

The affine transformation parameters are estimated by maximizing the Baum-Welch auxiliary function [12]. Denote

$$S_{\mathbf{O}}(m) = \sum_s \sum_t \gamma_{s,m}(t)\mathbf{o}_t,$$

$$S_M(m) = \sum_s \sum_t \gamma_{s,m}(t)\mu(s),$$

$$T(m) = \sum_s \sum_t \gamma_{s,m}(t),$$

$$S_{MM}(m) = \sum_s \sum_t \gamma_{s,m}(t)\mu(s)\mu^T(s),$$

$$S_{\mathbf{O}M}(m) = \sum_s \sum_t \gamma_{s,m}(t)\mathbf{o}_t\mu^T(s)$$

where $\mathbf{o}(t)$ is the observation at time t, $\gamma_{s,m}(t)$ is the the *a posteriori* probability of occupying HMM state s and Markov chain state m at time t given that the observation sequence $\mathbf{O} \equiv \mathbf{o}_1 \cdots \mathbf{o}_T$ is generated, and let

$$S_1(m) \equiv S_{MM}(m) - \tfrac{1}{T(m)} S_M(m) S_M^T(m),$$

$$S_2(m) \equiv S_{\mathbf{O}M}(m) - \tfrac{1}{T(m)} S_{\mathbf{O}}(m) S_M^T(m),$$

the parameters of the affine transform corresponding to state m of the Markov chain are represented by

$$\mathbf{A}(m) = S_2(m) S_1^{-1}(m)$$

and

$$\mathbf{B}(m) = \frac{1}{T(m)}[S_{\mathbf{O}}(m) - \mathbf{A}(m) S_M(m)],$$

where $m = 1, 2, \cdots, M$.

III. Experiments

The algorithm is tested on the Nov92 development database of WSJF0 which contains 16, 14, 17 and 14 sentences for four female speakers spk050, spk053, spk420 and spk421, respectively. Since the word

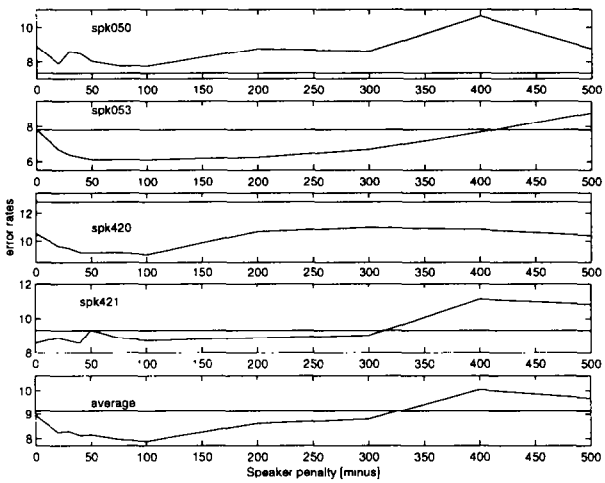

Fig. 1. Four-speaker error rates. Horizontal lines represent error rates obtained by using a speaker-independent model. Curves represent error rates in the multi-speaker speech recognition approach.

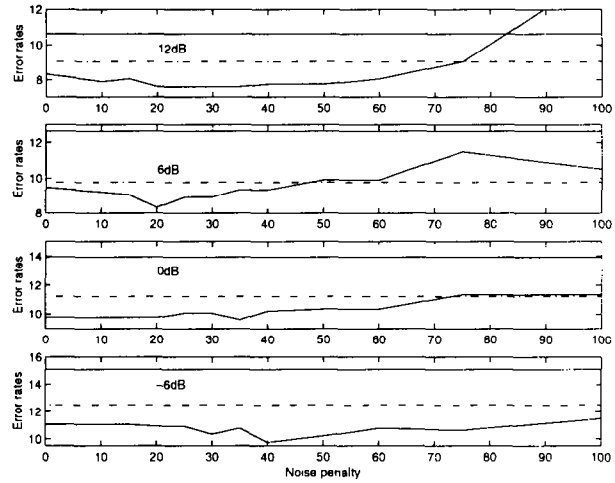

Fig. 2. Error rates for spk050 under machinegun noise. Horizontal solid lines are error rates obtained by using the speaker-independent model for clean speech. Horizontal dashdotted lines indicate error rates when the noise Markov chain has one state and curves indicate error rates when the noise Markov chain has two states.

graph is given in the second pass experiment, even if the acoustic phonetic score in the second pass is set to 1 for each frame, we can also get a word accuracy of 83.5% which can be regarded as the base-line score.

A. Four-speaker speech recognition

In four-speaker speech recognition, we study the recognition of speech spoken alternately by four speakers by using NA. The ergodic Markov chain has four states and the penalty is called the *speaker penalty.*

Fig. 1 shows the word error rate of each speaker and their average as a function of the speaker penalty. We find that the average word error rate is 9.1% when a speaker-independent model is used. However, when the speaker-independent model is adapted by NA and the speaker penalty is set to about -100, we get the minimal word error rate for each speaker and the corresponding average word error rate is 7.9%, a 13.6% improvement with respect to that obtained using the speaker-independent model.

The error rate of spk050 in the multi-speaker speech recognition strategy is higher than that obtained using the speaker-independent model for all values of the speaker penalty. This shows that NA cannot guarantee recognition improvement for each speaker, especially for those speakers whose error rate is already very low when using the speaker-independent model. This is caused by speaker misidentification due to the NA algorithm.

B. Speech corrupted by machinegun noise

When speech is corrupted by machinegun noise, the noisy speech is described by an HMM for clean speech and an ergodic Markov chain for the noise and the penalty is called the *noise penalty.* It is natural to describe the machinegun noise as a Markov chain with two states – one state for silences and the other for bursts. On the other hand, if the machinegun noise is treated as stationary noise, the corresponding Markov chain has only one state. In this case, there will be no inter-state transition, and hence no noise penalty.

Fig. 2 shows error rates of spk050 when her sentences are corrupted by additive machinegun noise at SNRs of 12 dB, 6 dB, 0 dB and -6 dB respectively. It is found from Fig. 2 that even if the machinegun noise is regarded as stationary noise and is represented by a one-state Markov chain, error rates can be greatly decreased. However, when the machinegun noise is represented by a Markov chain of two states, the error rate strongly depends on the noise penalty. It is consistently found that when the noise penalty is between 20 and 40, the error rates reach their minimal value which, compared with the results of a one-state Markov chain, lead to further substantial decrease. Table I lists error rate improvements by using NA. When the machinegun noise is treated as a stationary noise (one-state Markov chain), the average improvement is around 18%. When it is represented by a two-state Markov chain, the average improvement is around 32%. This shows the nonstationary characteristics of machinegun noise and the effectiveness of NA in the recognition of speech under nonstationary adverse environments.

C. Cross-talk problem

As a preliminary step for the application of NA to the cross-talk problem, the experiment is designed with the following assumptions: (i) The background

SNR	improvement	
	one state	two states
12 dB	14.9%	28.4%
6 dB	22.7%	34.1%
0 dB	19.6%	30.9%
-6 dB	17.2%	35.3%
Average	18.0%	32.0%

TABLE I

ERROR RATE IMPROVEMENT BY SUING A ONE- AND A TWO-STATE NOISE MARKOV CHAIN WHEN THE SPEECH IS CORRUPTED BY THE MACHINEGUN NOISE.

speech is spoken by a single speaker; (ii) The background speech is spoken by the same speaker in the testing and the adaptation data. Also we only consider the case where the SNR is 6 dB for the speech of all four speakers. When the speaker-independent model is used without adaptation, the average word error rate is 20.2%. However, even if we regard the background speech as stationary noise, in which case the Markov chain has only one state, we found that the average word error rate is decreased by 16.8%.

When a multi-state Markov chain is used, the way the phonemes are clustered into different states is crucial to the final recognition rate. We found that the best two-state partition is to cluster all vowels and consonants into one state and the silence itself into the other state. The average word error rate is decreased by 22.0%.

If a three-state Markov chain is used, a series of experiments showed that the best three-state partition of all the phonemes is to cluster all vowels into one state, all consonants into another state, and the silence itself into the third state. The error rate is decreased by 24.4%. Figure 3 represents the error rates as a function the noise penalty in the cross-talk problem. When we further differentiate these phoneme classes into more states, little improvement is obtained in recognition rate.

IV. Conclusion

In this paper, we have introduced a new approach, nonstationary adaptation, for speech recognition in nonstationary adverse environments. While greatly decreasing the word error rates in various nonstationary adverse environments, the computational cost of our algorithm is reasonable and the memory requirements are almost the same as those of clean speech recognition.

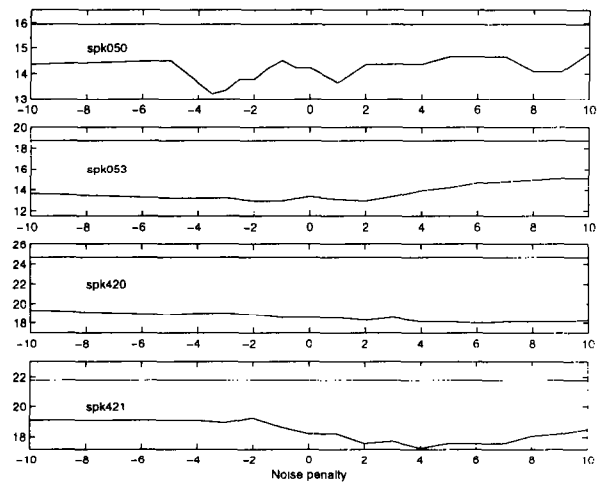

Fig. 3. Error rates in the cross-talk problem. Horizontal lines are error rates by using the speaker-independent model for clean speech. Curves represent error rates obtained by using NA and the Markov chain has three states.

REFERENCES

[1] R. Cardin, Y. Normandin and E. Millie, "Inter-word coarticulation modeling and MMIE training for improved connected digit recognition", *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, Vol. II, pp. 243-246, 1993.

[2] S. Das, R. Bakis, A. Nadas, D. Nahamoo and M. Picheny, "Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system", *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, Vol. II, pp. 71-74, 1993.

[3] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtraction (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars", *Speech Communication*, Vol. 11, Nos. 2-3, pp. 215-228, 1992.

[4] H. Hermansky, (1990), "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, Vol. 87, pp. 1738-1752, April, 1990.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. ASSP*, Vol. 27, No. 2, 1979.

[6] K. Shikano and F. Itakura, "Spectrum distance measures for speech recognition", in S. Furui and M. Sondhi, editors, *Advances in speech signal processing*, pp. 419-452, Marcel Dekker, Inc., 1991.

[7] A. P. Varga, R. K. Moore, "Hidden Markov Model decomposition of speech and noise", *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, pp. 845-848, 1990.

[8] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination", *Computer Speech and Language*, Vol. 9, No. 4, 1995.

[9] P. C. Woodland, M. J. F. Gales and D. Pye, "Improving environmental robustness in large vocabulary speech recognition", *Proc. of the International Conference on Acoustics, Speech and Signal Processing* Vol. I, pp. 65-68, 1996.

[10] C. Legetter, P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, Vol. 9, No. 2, 1995.

[11] J. L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "Developments in continuous speech dictation using the 1995 ARPA NAB news task", *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, 1996, Vol. I, pp. 73-76, 1996.

[12] L. E. Baum and J. A. Egon, "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology", *Bull. Amer. Meteorol. Soc.*, 73, pp. 306-363, 1967.

[13] A. P. Varga, H. J. M. Steeneken, M. Tomlinson and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition", Technical report, DRA Speech Research Unit, 1992.