PERFORMANCE EVALUATION OF ADAPTED AND RETRAINED MODELS FOR NOISY SPEECH RECOGNITION

HUANG Chao-Shih [1,2] *and Detlev Langmann* [1]

[1] Philips Innovation Center, Taipei (PICT)

[2] Department of Electrical Engineering, National Hsing-Hua University, Hsinchu

E-mail: {joseph, langmann}@prlt.research.philips.com

ABSTRACT

In this paper, experiments were performed to evaluate the principal performance boundaries of adapted and retrained models under added noise conditions. Adapted models fully alter the parameters of Hidden Markov Models (HMM) from clean ones in order to match the noisy test environment. Retrained models are fully trained from white Gaussian-noise contamined speech at matched signal-to-noise ratio (SNR) environments. We studied the capabilities and limitations of adapted models and models. The results show that retrained models perform better than adapted models under any conditions but especially for low SNRs. The results show that phone error rates for retrained models are about 6% better than for adapted models. It has also been found that the retrained models improve the word error rate by 6% for 15-dB SNR and even by 18% for 0-dB SNR.

1 Overview

The mismatch between training and testing conditions often causes a dramatic performance degradation, especially in real-world applications of automatic speech recognition (ASR) systems. Therefore, robustness algorithms become an important issue. The existing techniques can be grouped into three categories [1].

(i) Robust speech feature extraction: These techniques are based on noise-resistant features or robust distance measures without any noise estimation. It is difficult to find feature sets dealing with all types of noises. Short-term modified coherence (SMC) [2] has been found to be more robust against additive noise in small-vocabulary systems. More recently, one-side auto-correlation linear predictive coding (OSALPC)[3] has been found to obtain better performance than SMC in car environments. It has been reported that an integrated mel-scale with linear discriminant analysis (IMELDA) [4] could get more robust than standard mel-scaled cepstral coefficients (MFCCs). Perceptual linear predictive (PLP) analysis [5] and RelAtive SpecTrAl (RASTA) [6] processing perform better than LPC or MFCC parameters in noisy environment but they show degradations under clean conditions.

(ii) Speech enhancement: These techniques are to improve the SNR of corrupted speech and try to extract the clean speech from corrupted speech. Noise information is required that may come from a-priori statistics or online speech/pause detection. The enhanced speech although less noisy may contain a different form of distortion which could be more difficult to deal with. The simplest technique for clean speech estimation is to use spectral subtraction [7]. But, to avoid the negative spectral value a non-linearity must be introduced. Probabilistic optimal filtering (POF) [8] is mapped by a piece-wise linear transformation but the environment should be learnt. Algorithms based on codebook dependent cepstral normalization (CDCN) [9] attempt to learn a mapping from corrupted speech to clean speech, but some of them need stereo recording.

(iii) Model adaptation/ transformation: These techniques adapt the models onto the test environments. The resulting performance depends on the state-to-frame alignment and is often bounded by the performance in matched conditions. Maximum likelihood linear regression (MLLR) [10] is used to adapt the models into new environments. Stochastic matching (SM) [11] modifies features or models in order to match the environmental change. SM has much potential both for additive and convolutional noise. The speech and noise decomposition (SND) [12] and parallel model combination (PMC) [13] [14] is to model the background noise, instead of using a simple mean or relative value.

2 Simulation of Adapted and Retrained HMMs

We examined the principal limitations of algorithms of category (iii) by using fully adapted and retrained models. The fully adapted model is used to simulate that the added noise can be estimated accurately for model re-estimation. The experimental set-up of the fully adapted models is as follows:

Step1: The clean speech of training corpus is segmented by means of clean models, and the paths are kept for noisy model training.

Step2: Different levels of added noise are added into the test utterances. All HMM parameters are re-estimated without any further iteration.

The retrained models are fully trained from noisy speech at matched SNR environments like the training of clean models .

White Gaussian noise was added to the testing utterances at different total SNR levels. The total SNR is defined as follows, where σ_x^2 is the variance of testing speech utterance and σ_n^2 is the variance of added noise.

$$TotalSNR = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_n^2}\right) \quad \text{(dB)}$$

3 Experiments

The experiments were performed on the “Japanese Electronic Industry Development Association's Common Speech Data Corpus” (JSDC) being mainly an isolated-phrase corpus [15]. The JSDC corpus was recorded with dynamic microphones and sampled at 16 kHz. The phonetically rich JSDC city-name subcorpus was used to train phone-based HMMs. We deployed 35 monophone HMMs with three states per model and nominal 32 Laplacian mixture densities per state in our experiments. The JSDC control-word corpus with a vocabulary of 63 words was used as testing material. Experiments for free-phone decoding and word recognition were performed. The resulting phone and word error rates are shown in Fig.1 and Fig. 2, respectively.

1) Corrupted performance: The models are clean and the test material is corrupted by added white Gaussian noise, where clean means there is no noise added.
2) Fully adapted performance: The models are adapted from clean ones based on known noise levels and the test material is corrupted at the same SNR levels.
3) Retrained performance: The models are fully retrained in known SNR environments and the test material is corrupted at the same SNR levels.

We found that retrained models perform always better than adapted models under any condition but especially at low SNR levels. Fig.1 shows that phone error rates for retrained models are about 6% better than for adapted models. From Fig. 2, it also can be seen that retrained models improve the word error rate by 6% of for 15-dB SNR and even by 18% for 0-dB SNR.

4 Conclusion

We have shown that retrained models always provide better performance than fully adapted models when the SNR decreases. Retrained models are currently the possible solution for lower SNR levels. Our future work will focus on whether it is possible to obtain or to overcome the performance of retrained models in noisy environments.

Acknowledgements

The authors would like to express their sincere gratitude to Dr. LEE Chin-Hui for his constantly guidance and encouragement. Special thanks are also to Professor WANG Hsiao-Chuan for his supervision of Chao-Shih’s research. The authors also like to thank Dr. Terry Doyle and Dr. Paul Lin from PICT for their kind support for this work.

REFERENCES

[1] Y. Gong, “Speech Recognition in Noisy Environments: A Survey”, Speech Communication, Vol-16, pp261-291, 1995.

[2] D. Mansour and B. H. Juang, “ The Short-time Modified Coherence Representation and Noisy Speech Recognition”, IEEE Trans. ASSP, Vol-37, pp795-804, 1989.

[3] J. Hernando and C. Nadeu, “ Speech Recognition in Noisy Car Environment based on OSALPC Representation and Robust Similarity Measuring Techniques”, ICASSP, pp69-72, 1994.

[4] M. J. Hunt and C. Lefebre, “A Comparison of Several Acoustic Representation for Speech Recognition with Degraded and Undegraded speech”, ICASSP, pp262-265, 1989.

[5] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis of Speech”, JASA, Vol-87, pp1738-1752, 1990.

[6] J. Koehler etc., “Integrating RASTA-PLP into Speech Recognition”, ICASSP, pp421-424, 1994.

[7] J. S. Lim and A. V. Oppenhein, “ Enhancement and Bandwidth Compression of Noisy Speech”, IEEE Proceeding, Vol-67, pp1586-1604, 1979.

[8] L. Neumeyer and M. Weintraub, “Probabilistic Optimal Filtering for Robust Speech Recognition”, ICASSP, pp417-420, 1994.

[9] A. Acero and R. M. Stern, “Environmental Robustness in Automatic Speech Recognition”, ICASSP, pp849-852, 1990.

[10] C. J. Legetter and P. C. Woodland, “ Flexible Speaker Adaptation using Maximum Likelihood Linear Regression”, ARPA Workshop on Spoken Language System Technology, pp110-115, 1995.

[11] A. Sankar and C. H. Lee, “ Robust Speech Recognition based on Stochastic Matching”, ICASSP, pp121-124, 1995.

[12] A. P. Varga and R. K. Moore, “Hidden Markov Model Decomposition of Speech and Noise”, pp845-848, 1990.

[13] M.J.F. Gales and S. Y. Young, “ HMM Recognition in Noise Using Parallel Model Combination”, Prof. of Eurospeech, pp837-840, 1993.

[14] C. S. Huang and E. F. Huang, “Parallel Model Combination (PMC) for Robust Speech Recognition”, 1997 Workshop on Distributed System Technologies & Application, pp203-206, 1997.

[15] J. Picone, J. Hamaker , and R.J. Duncan, " JEIDA Corpus of Japanese Common Speech Data", Mississippi State Unversiy, 1986.

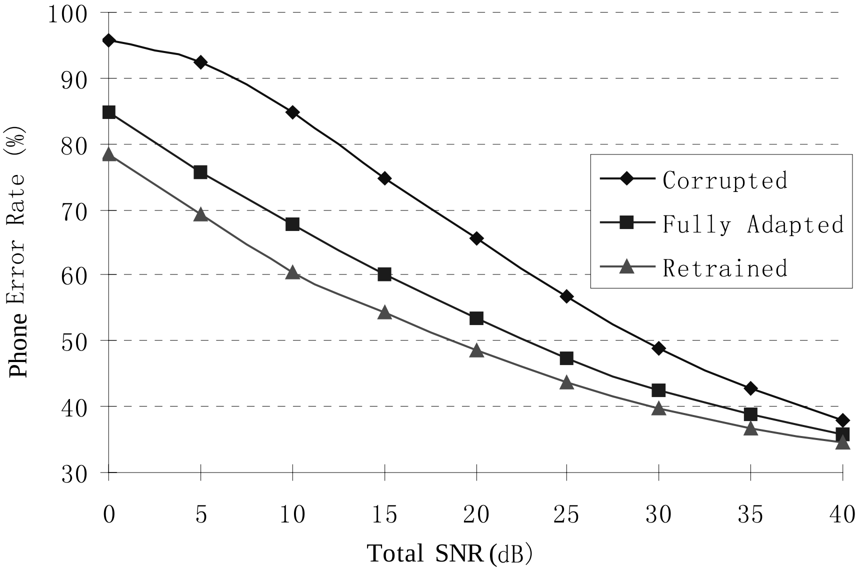

Fig.1. Comparisons of free phone decoding performance for corrupted, fully adapted, and retrained models.

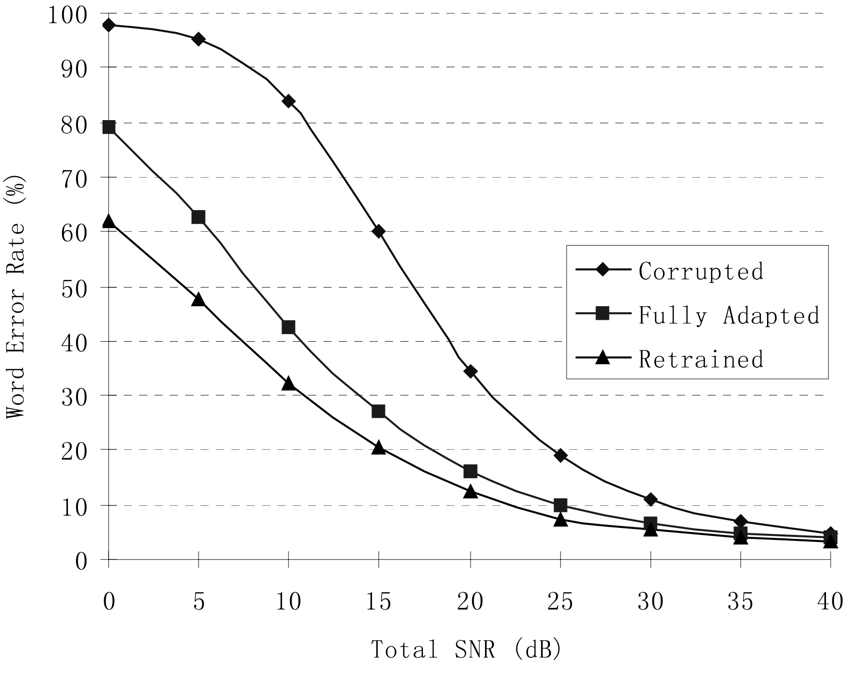

Fig.2. Comparisons of word recognition performance for corrupted, fully adapted, and retrained models.