UNSUPERVISED HMM ADAPTATION BASED ON SPEECH-SILENCE DISCRIMINATION

Ilija Zeljković, Shrikanth Narayanan and Alexandros Potamianos

AT&T Labs – Research, 180 Park Avenue, P.O. Box 971, Florham Park NJ 07932
email: {ilija,shri,potam}@research.att.com

ABSTRACT

An unsupervised, sentence-level, discriminative, HMM adaptation algorithm based on silence-speech classification is presented. Silence and speech regions are determined either using an endpointer or using the segmentation obtained from the recognizer in a first pass. A unsupervised discriminative training procedure using the gradient descent algorithm, with N-best competing strings with word insertions is then used to improve the discrimination between silence and speech. Experiments on connected digits show about 40-80 % reduction in insertion errors, a small amount of reduction in substitution errors, and a negligible effect on deletion errors. In addition, experiments on noisy speech showed about 70% word error rate reduction, thus demonstrating the robustness of the proposed adaptation technique.

1. INTRODUCTION

Acoustic mismatch between training and testing conditions results in significant performance degradation in hidden Markov model (HMM)-based speech recognizers. Careful inspection of the recognition errors shows that word insertion and substitution errors often occur as a result of poor recognition scores for acoustic segments with low energy. Low-energy portions of the speech signal tend to be highly confusable with silence, especially when channel and noise mismatch exists between the speech signal and the acoustic models. Various blind deconvolution and bias removal schemes (for e.g., cepstral mean normalization [1]) exist that address this problem of mismatch at the utterance level. This paper focuses on reducing the utterance-model mismatch at those segments of the speech signal where the acoustic characteristics of the background and the speech signal, typically unvoiced portions, are similar.

Our goal in this paper is to adapt the parameters of the acoustic model, in an unsupervised mode, during the recognition process in order to improve discrimination between the background model and the speech models. One ("instantaneous adaptation") or more ("long-term adaptation") utterances can be used for adaptation. The main idea is to increase the separation between the *correct* string and *competing* string candidates. Although the correct string is not known during the recognition process, this study exploits the fact that the silence regions of the sentence can be determined with greater accuracy and efficiency and these regions can be used for discriminative training.

The organization of this paper is as follows. First we present an overview of the relevant literature followed by the description of our proposed speech-silence discrimination adaptation algorithm. In Section 2.1, implementation issues are discussed. Specifically, novel ways of producing competing recognition hypothesis for the adaptation algorithm are proposed. Finally, in Section 3, we apply the algorithm to a connected digit recognition task and show digit error rate reduction up to 70-80%.

1.1. PREVIOUS WORK

A significant part of the speech recognition literature deals with problems caused to real-world recognition systems by noise, distortion or variability in the speech waveform. Various algorithms such as cepstral mean normalization, maximum likelihood (ML) cepstrum bias normalization [1], ML frequency warping [2], and ML linear regression [3] have been proposed to deal with these problems. Apart from these transformation-based techniques that produce good results with a limited amount of adaptation data, the acoustic models can be retrained using Bayesian adaptation. Bayesian adaptation typically requires a large amount of adaptation data; algorithms have been proposed for updating groups of HMM parameters or for smoothing the re-estimated parameter values, for e.g., vector field smoothing, classification tree or state-based clustering of distributions [4]. Parallel model combination (PMC) has been also used to combat both additive noise distortion and multiplicative (channel) distortion [5].

Typically, these algorithms perform well for simulated data, i.e., when additive or multiplicative distortion is added to the speech signal in the laboratory but not equally well in field trials where a multitude of sources with time-varying characteristics can distort the speech signal. In many cases, very little data are available for adaptation. Further, the adaptation data might not be transcribed. For example, discriminative training of HMMs, which helps improve recognition accuracy [8, 6, 7], assumes that the linguistic context of the utterance is known. Unsupervised adaptation using very few utterances is a difficult problem because there are no guarantees that the adapted parameters will converge toward global optimum values. Utterance verification algorithms can be used to make unsupervised adaptation more robust.

In this paper, we focus on the problem of adapting the speech and silence acoustic models to improve our ability to discriminate between speech and silence regions. In the next section, we propose novel ways of producing competing recognition hypotheses. Unsupervised discriminative training with generalized probabilistic descent algorithm (GPD) is used to adapt the speech and silence acoustic models [9].

2. SPEECH-SILENCE ADAPTATION ALGORITHM

The short-term and long-term spectral characteristics of the background part of speech utterances collected through the public switched telephone network are highly variable. Although HMM-based recognizers often confuse the background with valid speech segments, thereby producing insertion or deletion errors, portions of the background (*silence*) regions can be identified with fairly high accuracy using simpler but more robust techniques using features such as the short-time energy and the short-time zero-crossings. The high degree of certainty in determining *silence* regions and the use of only these regions to adapt both speech and silence models makes the speech-silence adaptation algorithm both accurate and efficient.

The adaptation algorithm consists of the following four steps:

1. Split the input utterance into *speech* and *silence* regions. If more than one background HMM is used, do optimal (Viterbi) decoding of *silence* regions using the background HMMs.
2. Generate *competing* strings by aligning the states of speech HMM to *silence* regions (i.e., artificially encourage or force insertions).
3. Enhance separation between HMMs in *correct* and *competing* strings using a discriminative training algorithm.
4. Perform optimal decoding (recognition) on the whole utterance using the newly adapted HMMs and any prescribed grammar.

There are many ways to implement steps 1,2 and 3 of the algorithm proposed above. Next, we discuss a simple and efficient implementation of the adaptation algorithm.

2.1. IMPLEMENTATION ISSUES

Speech-silence segmentation (step 1) may be obtained by a simple preprocessing step before the recognition process begins. In the current implementation the silence-speech segmentation is performed by the recognizer in the first pass using the initial HMMs, and a grammar, with no insertion penalties. This is assumed to be the "correct string". Competing strings (step 2) are produced in two alternative ways:

(a) *Acoustically-driven insertion*: A negative insertion penalty (insertion *incentive*) is used to decode four best competing strings (encouraged internal insertion).

(b) *Blind external insertion*: Eleven competing strings (for digit recognition test) are generated: each digit is added before and after the initially recognized string, generating one competing string (forced external insertion).

The discriminative training [7, 6] (step 3) is performed by using the minimum string-error training algorithm using N competing string models [9]. A brief description of the discriminative training algorithm using GPD is given in the next section (Sec. 2.2). Finally the second-pass recognition is performed with the adapted models using the Viterbi decoding algorithm (step 4).

2.2. DISCRIMINATIVE TRAINING

The goal of the discriminative model training algorithm [7, 6] is to find a model set that optimally distinguishes between observation sequences corresponding to correct class models and those of N

competing class models by maximizing the mutual information between the observation sequence O and the words or strings of that class (represented by a parametrized HMM, Λ). The misclassification measure

$$d(O,\Lambda) = -g(O,S_\nu,\Lambda) + \log\{\frac{1}{N-1}\sum_{S_k \neq S_\nu} e^{g(O,S_k,\Lambda)\eta}\}^{\frac{1}{\eta}} \tag{1}$$

uses the discriminant function

$$g(O,S_k,\Lambda) = \log f(O,\Theta_{S_k},S_k \mid \Lambda) \tag{2}$$

which is defined in terms of the loglikelihood score f on the optimal state sequence Θ_{S_k} (given the model set Λ) for the k^{th} best string, S_k. The discriminant function for the transcribed training string S_ν is $g(O,S_\nu,\Lambda)$. The model loss function for string error rate minimization, $l(O,\Lambda) = 1/(1+exp(-\gamma d(O,\Lambda))$, where γ is a positive constant, is solved using gradient descent algorithm [8, 9].

As will be shown in the next section, the N competing strings can be generated directly from the acoustics or externally by some simple blind string appending scheme.

3. RECOGNITION EXPERIMENTS

Speech units (words and subwords) as well as background silence are modeled by first order, left-to-right HMMs with continuous observation densities. The observation vector consists of 39 features: 12 LPC derived *cepstral* coefficients, dynamically normalized energy, as well as their first and second derivatives. Eleven digits, including "*oh*" and "*zero*", were used in the evaluation task. Each digit was modeled with either 20 or 15 state HMMs, with 16 Gaussian mixtures. Speech background (*silence*) is modeled with a single state, 128 Gaussian mixture HMM. The HMMs were trained using data extracted from speech data collected over the telephone network (16089 digit strings).

In the recognition process, the sequence of observation vectors from an unknown speech utterance are matched against a set of stored hidden Markov models representing speech units. A search network is generated by a finite state grammar that describes the set of valid strings. The network search algorithm returns the single most likely sequence of speech units. The search procedure is a Dynamic Programming (DP) algorithm (Viterbi decoding) where the goal is to find a valid state sequence with the highest accumulated state log-likelihood.

Experimental Results: The algorithm was tested on speech data collected from two AT&T service trials. Trial I data, consisting of 10768 16-digit strings, represented *matched* training and testing conditions. On the other hand, no data from Trial II were represented in training. Moreover, Trial II data consist only of single digits (a total of 2159 utterances). It should be pointed out that isolated digits represented only a small portion of the training database. In addition, a test set with noisy speech data (601 digit strings spoken over the telephone network), marked by human listeners, was used to verify the robustness of the proposed algorithm.

Tables 1 and 2 summarize the recognition results for various testing conditions: Results are compared for the two methods of competing string generation (N-best competing strings by acoustically-driven insertion using insertion incentives and blind external insertion by forced initial and final digit appending), with each case repeated with and without resetting the models to the baseline class for each new string input. The baseline results correspond to no model adaptation.

Under reasonably matched training and testing conditions, we observe that insertion errors are reduced in all test cases when adaptation is used. The best results are obtained for the case that uses competing strings generated through insertion incentives. Moreover, as expected, long-term adaptation (using all available utterances for adaptation) performs better than instantaneous adaptation (i.e., a single utterance is used to adapt the HMMs). On the other hand, although the blind insertion method has a similar effect on insertion errors, it is accompanied by increased substitution and deletion errors, particularly in the long-term adaptation case, suggesting divergence in the adapted models with increasing adaptation data.

The unusually high number of insertion errors in the baseline results for Trial II data is attributed to the structural mismatch between the training data and this particular test set which is composed entirely of isolated digits. Instantaneous adaptation gives about 36-38% improvement in word error rates for both methods of competing string generation. For long-term adaptation, however, the blind insertion method of competing string generation yields poorer performance than the baseline while the acoustically-driven insertion method yields more than 80% improvement in word error rate. A closer analysis of the results shows that although there is improvement in insertion errors (which is indeed the objective of our proposed algorithm), there is significant increase in substitution errors for the blind insertion method. This result further supports our earlier remark that model divergence (instability) with increasing adaptation data is a potential pitfall when blind insertion is used for competing string gener-

		Word Error (%)			
Competing string generation	Adaptation mode	Total	Sub	Del	Ins
None (baseline)	N/A	1.25	0.8	0.1	0.4
acous. driven ins	long-term	1.08*	0.8	0.1	0.2
blind ins	long-term	1.23	0.9	0.2	0.2
acous. driven ins	instant.	1.16	0.7	0.1	0.3
blind ins	instant.	1.17	0.7	0.1	0.3

Table 1: Recognition performance on Trial I test data (N = 10768 strings) with HMM adaptation: *matched* training and testing conditions. Results are shown for instantaneous and long-term adaptation and for the two methods of competing string generation: acoustically-driven by insertion incentives and blind insertion by word appending.

		Word Error (%)			
Competing string generation	Adaptation mode	Total	Sub	Del	Ins
None (baseline)	N/A	12.20	1.4	0.0	10.8
acous. driven ins	long-term	2.11*	1.3	0.0	0.8
blind ins	long-term	15.3	12.5	0.0	2.9
acous. driven ins	instant.	7.56	1.3	0.0	6.1
blind ins	instant.	7.8	1.3	0.0	6.4

Table 2: Recognition performance on Trial II test data (N = 2159) with HMM adaptation: *mismatched* testing and training conditions. Other testing conditions similar to those for Table 1.

ation.

Table 3 shows the performance results on the noisy speech data. Baseline results show high insertion error rates which decrease dramatically with the new adaptation strategy. Although there is slight increase in the deletion errors, the overall error rate improvement is dominated by the decrease in the insertions.

Recognition details	Word Error %	Sub %	Del %	Ins %
Baseline	15.9	13.8	0.2	1.9
blind ins	5.1	2.3	0.9	1.9
acous. driven ins	4.9	2.2	0.9	1.8

Table 3: Recognition performance on digit strings (N = 601) with appreciable background noise (as marked by human listeners) with instantaneous HMM adaptation.

4. SUMMARY

A novel HMM adaptation method based on speech-silence discrimination was presented. In summary, the main contributions of this work are:

- The exclusive use of signal portions declared by the algorithm as *silence* segments (i.e., unsupervised modality) for adapting both silence and some/all speech models in a way that results in improved speech-silence discrimination in the new model set.
- Automatic competing string generation by providing insertion incentives, inserting words that are naturally prone to acoustic confusion with background.
- Unsupervised adaptation using the gradient descent algorithm that assures convergence.

Results show that competing strings directly provided by the recognizer by employing insertion incentives give the most useful set of data for speech-silence discrimination, and yields the best overall error rate improvements even under mismatched training and testing conditions. Dramatic improvements in insertion errors were obtained for telephone speech data with high background noise.

5. REFERENCES

[1] A. Sankar and C.-H. Lee, "Robust speech recognition based on stochastic matching," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, pp. 121–124, May 1995.

[2] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, pp. 353–356, May 1996.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[4] J. Ishii, M. Tonomura, and S. Matsunaga, "Speaker adaptation using tree structured shared-state HMMs," in *Internat. Conf. Speech Language Processing*, Oct. 1996.

[5] Y. Minami and S. Furui, "Adaptation method based on HMM composition and EM algorithm," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, pp. 327-331, May 1996.

[6] Normandin Y., Cardin R, and De Mori R., "High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation", *IEEE Trans. on Speech and Audio Processing, Vol 2, no. 2, April 1994*

[7] Juang B. H., and Katagiri S., "Discriminative learning for minimum error rate training", *IEEE Trans. on Signal Processing, Vol 40, pp. 3043-3054, April 1994*

[8] Euler S., Zinke J., "Experiments on the Use of the Generalized Probabilistic Descent Method in Speech Recognition", *Proc. ICSLP, pp. 157-160, 1992*

[9] Chou, W., Lee C-H., and Juang, B-H., "Minimum Error Rate Training Based on N-best String Models", *Proc. ICASSP 1993, Vol 2, pp. 652-665.*

[10] Chou W., Rahim M., and Buhrke E., "Signal conditioned minimum error rate training", Proceedings of Eurospeech, pp. 495-499, Sept. 1995.