TRANSFORMING HMMS FOR SPEAKER-INDEPENDENT HANDS-FREE SPEECH RECOGNITION IN THE CAR

Y. Gong and John J. Godfrey

Speech Research, Media Technologies Laboratory, TEXAS INSTRUMENTS
P.O.BOX 655303 MS 8374, Dallas TX 75265, U.S.A.
{Yifan.Gong,Jack.Godfrey}@ti.com

ABSTRACT

In the absence of HMMs trained with speech collected in the target environment, one may use HMMs trained with a large amount of speech collected in another recording condition (e.g., quiet office, with high quality microphone). However, this may result in poor performance because of the mismatch between the two acoustic conditions.

We propose a linear regression-based model adaptation procedure to reduce such a mismatch. With some adaptation utterances collected for the target environment, the procedure transforms the HMMs trained in a quiet condition to maximize the likelihood of observing the adaptation utterances. The transformation must be designed to maintain speaker-independence of the HMM.

Our speaker-independent test results show that with this procedure about 1% digit error rate can be achieved for hands-free recognition, using target environment speech from only 20 speakers.

1. INTRODUCTION

Speech recognition for matched conditions (i.e., where training and testing are performed in the same acoustic conditions) has achieved low recognition errors, for instance, 1% word error rate (WER) for connected digits recorded over the telephone network [5]. However, such results are based on the collection of large amounts of training data under conditions as close as possible to the testing data. In many speech recognition tasks (e.g., hands-free recognition in a car), collection of a large database to train speaker-independent HMMs is very expensive. Yet if HMMs are used in cross-condition recognition, such as using a close-talking microphone in a quiet office for training, and then testing on hands-free recognition in a car, the mismatch will degrade recognition performance substantially.

In terms of power spectral density, the mismatch can be characterized by a linear filter and an additive noise: $|Y(\omega)| = |H(\omega)|^2 \cdot |X(\omega)| + |N(\omega)|$ where $Y(\omega)$ represents the speech to be recognized, $H(\omega)$ the linear filter, $X(\omega)$ the training speech, and $N(\omega)$ the noise. In the log spectral domain, this equation can be written as:

$$\log|Y(\omega)| = \log|X(\omega)| + \psi(N(\omega), X(\omega), H(\omega)) \quad (1)$$

with

$$\psi(N(\omega), X(\omega), H(\omega)) \triangleq \log|H(\omega)|^2 + \log\left(1 + \frac{|N(\omega)|}{|X(\omega)| \cdot |H(\omega)|^2}\right). \quad (2)$$

ψ can be used to characterize the mismatch, which depends on the linear filter, the noise source and the signal itself.

To overcome the mismatch, several types of solutions have been reported. For example, cepstral mean normalization (CMN) is known for its ability to remove the first term in ψ (i.e., stationary bias) in cepstra [2]. –¿ For example, cepstral mean normalization (CMN) removes stationary bias by removing the first term in ψ. ¡– It has been shown that using CMN, telephone quality speech models can be trained with high quality speech [9]. However, this is not effective for the second term, which is caused by additive noise and cannot be assumed constant within the utterance. Two-level CMN [4] alleviates this problem by introducing a speech mean vector and a background mean vector. Other, more detailed models of the mismatch include joint additive and convolutive bias compensation [1] and channel and noise estimation [8].

In this paper, we assume that we are given two datasets:

- An initial set of HMMs trained on large amount of speech recorded in one condition (e.g., in a quiet

room using high quality microphone), which provides rich information on coarticulation and speaker variability, and

- A smaller speech database collected in the target environment, which provides information on the test condition including channel, microphone, background noise and reverberation.

We are interested in training accurate speaker-independent HMMs for use in the target environment.

We propose a model adaptation approach to obtain the target models. For this purpose, maximum likelihood linear regression (MLLR) [6] is adopted, which models any distortion between the initial HMM and the target environment as a set of state-dependent linear transformations. It has the potential of compensating for a combined effect of channel and background noises. MLLR has been very successful for speaker-dependent adaptation [6, 11]. However, reports on MLLR for speaker-independent HMM transformation for noisy speech recognition have not been found.

2. PROCEDURE

The procedure takes as input two sets of data: utterances collected in one environment (R) for a few hundred speakers, and utterances collected in the target application or environment (T) for some much smaller number of speakers. As output, the procedure gives a set of HMMs that are adjusted and suitable for recognizing further speech from the target application task. The new set of models is speaker independent, and the adjustment is a one-time operation for each application task.

We propose the following procedure:

1. train a set of HMMs H with data collected in environment R
2. repeat until the likelihood of data T stabilizes:
   2.1. transcribe phonetically all utterances in T using H
   2.2. group transcribed segments into a set of classes C
   2.3. $\forall c \in C$:
      2.3.1. find transformation Φ_c to maximize the likelihood of utterances in T
      2.3.2. transform HMMs using Φ_c: $H \leftarrow \Phi_c(H)$

The steps in 2. are explained as follows:

Step 2.1

Viterbi decoding is used to locate the starting and ending time frame of each phone segment.

Step 2.2

The phone segments are grouped into phonetic classes according to their acoustic similarity and the number of frames within a segment. The greater the number of frames, the larger the number of resulting phonetic classes.

Step 2.3.1

The transformation Φ_c of the class c changes the mean vector of the Gaussian distribution of HMMs according to:

$$\mu_{j,k,h} = \Phi_c \hat{\mu}_{j,k,h} \tag{3}$$

where $\mu_{j,k,h}$ is the transformed mean vector for state j, mixture component k of the HMM h, and $\hat{\mu}_{j,k,h}$ is the original mean vector, which has the form:

$$\hat{\mu} = [\omega, \mu_1, \mu_2, \ldots \mu_n]' \tag{4}$$

where ω is the offset of the regression.

The observation probability density of Gaussians mixture HMMs is in the form of:

$$b(o|j,k,h) = \frac{\exp(-\frac{1}{2}(o - \Phi_c\hat{\mu}_{j,k,h})'\Sigma^{-1}_{j,k,h}(o - \Phi_c\hat{\mu}_{j,k,h}))}{(2\pi)^{\frac{n}{2}}|\Sigma_{j,k,h}|^{\frac{1}{2}}} \tag{5}$$

Following [6], the transformation Φ_c that maximizes the likelihood is given by the following matrix equation:

$$\begin{aligned}\sum_{h\in c}\sum_{s\in S_h}\sum_{t\in T_s}\sum_{j\in\theta_h}\sum_{k\in\alpha_{h,j}} \gamma^{(s,t)}_{j,k,h}\Sigma^{-1}_{j,k,h}o^s_t\hat{\mu}'_{j,k,h} = \\ \sum_{h\in c}\sum_{s\in S_h}\sum_{t\in T_s}\sum_{j\in\theta_h}\sum_{k\in\alpha_{h,j}} \gamma^{(s,t)}_{j,k,h}\Sigma^{-1}_{j,k,h}\Phi_c\hat{\mu}_{j,k,h}\hat{\mu}'_{j,k,h}\end{aligned} \tag{6}$$

where:

- S_c is the set of all segments grouped into class c,
- T_s is the utterance frames in the segment s,
- θ_h is the states in the HMM h,
- $\alpha_{h,j}$ is all mixture components of HMM h at state j, and
- $\gamma^{(s,t)}_{j,k,h}$ is the probability of being in state j at time t with mixture component k, for the segment s of the model h.

Equation 6 represents a linear system of Φ_c and can be solved by any appropriate technique.

Step 2.3.2

In this step, all HMMs of the class c are updated using the new transformation.

3. EXPERIMENTAL RESULTS

3.1. Speech databases

The procedure is evaluated for connected digit recognition.
Initial model training materials:
The environment R consists of the TI-DIGITS database [7], down-sampled to 8 kHz. This database was recorded over a high quality microphone in a very quiet room. The training part includes 8614 utterances of from 1 to 7 digits by 112 speakers.
Adaptation materials:
The environment T consists of utterances, also sampled at 8 kHz, recorded inside a parked car over a hands-free microphone mounted on the visor. Twenty speakers each read 40 strings of 4, 7, or 10 digits.
Evaluation materials:
For the same 20 speakers, a separate recording session yielded 800 strings of 4, 7, or 10 digits. The number of digits in the utterance was kept unknown to the recognizer during the evaluation.

The observation vectors consist of 10 DFT mel-frequency cepstral coefficients (MFCC) along with their regression-based first-order time derivative at a frame rate of 20ms.

3.2. HMM structure

We use gender-dependent digit-specific phone models. There are about 78 HMMs per gender. Each state has self-loop, jump to next state and skip transitions. The number of states for a phone model is based on the average duration of phone segments in the training data, with some exceptions for practical considerations. Up to 8 Gaussian distributions per state are used. The HMMs are trained using an EM procedure.

3.3. Evaluation procedure and results

Since we want to know the performance of the speaker-independent model transformation, and our evaluation data set has only 20 speakers, a jack-knife procedure was used in order to ensure speaker-independence of the tests:

1. Repeat until all speakers are tested:
   1.1. select an untested speaker
   1.2. transform HMMs using the utterances of the remaining 19 speakers
   1.3. test the recognition performance on the selected speaker
2. Average the recognition results over speakers.

The recognizer was tested under four configurations:

- Direct use of HMM models trained in R for recognition of utterances in T. Preliminary tests show that this resulted in severe performance degradation due to microphone mismatch and background noise.
- One unique bias ($\mu = \hat{\mu} + B$) for all phone models. The WER drops to 1.54%. Note that unique bias should outperform CMN, since it makes use of HMM structure [10].
- With one unique linear transformation ($\mu = A\hat{\mu} + B$) for all phone models. The WER is reduced to 1.2%.
- With a number N of linear transformations (Eq-3), where N is controlled in the experiments by M, the minimum number of frames required to introduce a transformation class. Figure 1 shows the word error rate as a function of M. We can see that for $M = 500$, the lowest WER of 1.02% is obtained.

To summarize, if we take *unique bias* as the baseline, then *unique linear transformation* reduces the error rate by about 20% and the best set of *N linear transformations* reduces the error rate by about 35%.

3.4. Discussion

Figure 1 shows that:

- The transformation is not optimal if the minimum number of vectors per transformation is too large (e.g. $M = 1000$). This confirms that the mismatch caused by the combination of distortion factors is state-dependent and therefore needs to be covered by an adequately large number of state-dependent transformations. Actually, if $M = \infty$ then the system reduces to the case of *unique linear transformation* as the WER curve indicates.
- Too many transformations result in poor performance (as $M \rightarrow 0$), because in this case the transformations will capture information specific to the 19 adaptation speakers in addition to the acoustic conditions of the adaptation utterance.

Some test utterances contain unusually high background noise. This means that the WER could be further reduced if some noise resistant recognition feature, such as parallel model combination [3], were included.

4. CONCLUSION

We presented a procedure for obtaining accurate speaker-independent HMMs for hands-free digit recognition using initial HMMs trained on "high-quality" speech and using some adaptation utterances collected in the target environment.

The procedure uses state-dependent linear transformations which are adjusted to yield speaker-independent performance.

Experimental results show that the procedure gives substantial WER reduction over simple cepstral mean normalization. Our cross-environment recognition evaluation achieves performance (1% WER) similar to that of conventional environment-dependent, matched-condition training for similar tasks, such as digit recognition over the telephone network, yet only 20 speakers had to be collected in the target environment.

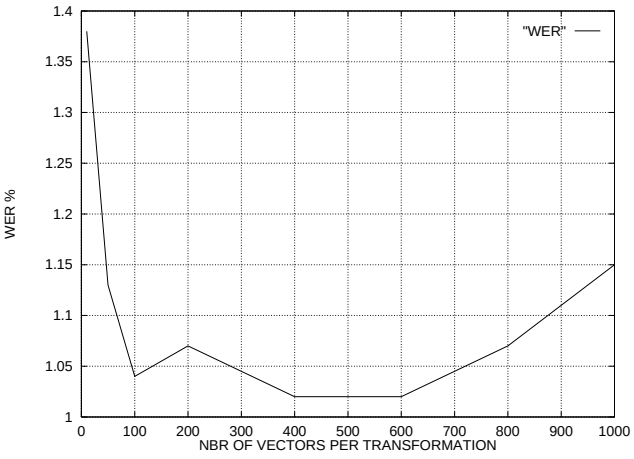

Figure 1: WER as function of minimum number of vectors per transformation

5. REFERENCES

[1] M. Afify, Y. Gong, and J.-P. Haton. A unified maximum likelihood approach to acoustic mismatch compensation: Application to noisy lombard speech recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Germany, 1997.

[2] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-29(2):254–272, April 1981.

[3] M. J. F. Gales and S. J. Young. HMM recognition in noise using parallel model combination. In *Proceedings of European Conference on Speech Communication and Technology*, volume II, pages 837–840, Berlin, 1993.

[4] S. K. Gupta, F. Soong, and R. Haimi-Cohen. High-accuracy connected digit recognition for moble applications. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, pages 57–60, Atlanta, May 1996.

[5] Y. H. Kao and L. Netsch. Inter-digit HMM connected digit recognition using the MACROPHONE corpus. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, Germany, 1997.

[6] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer, Speech and Language*, 9(2):171–185, 1995.

[7] R. G. Leonard. A database for speaker-independent digit recognition. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, pages 42.11.1–42.11.4, San Diego, 1984.

[8] D. Matrouf and J. L. Gauvain. Model compensation for noises in training and test data. In *Proc. of IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, Germany, 1997.

[9] L. G. Neumeyer, V. V. Gigalakis, and M. Weintraub. Training issues and channel equalization techniques for the construction of telephone acoustic models using a high-quality speech corpus. *IEEE Trans. on Speech and Audio Processing*, 2(4):590–597, October 1994.

[10] M. G. Rahim and B.-H. Juang. Signal bias removal by maximum likelihood estimation for bobust telephone speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4(1):19–30, Jan 1996.

[11] O. Siohan, Y. Gong, and J.-P. Haton. Comparative experiments of several adaptation approaches to noisy speech recognition using stochastic trajectory models. *Speech Communication*, 18:335–352, 1996.