SPEAKER ADAPTATION BY MODELING THE SPEAKER VARIATION IN A CONTINUOUS SPEECH RECOGNITION SYSTEM

Nikko Ström

Dept. of Speech, Music and Hearing, KTH, Sweden

ABSTRACT

A method for unsupervised instantaneous speaker adaptation is presented and evaluated on a continuous speech recognition task in a man-machine dialogue system. The method is based on modeling of the systematic speaker variation. The variation is modeled by a low-dimensional speaker space and the classification of speech segments is conditioned by the position in the speaker space. Because the effect of the speaker space position on the classification is determined in an off-line training procedure using the speakers in a training database, complex systematic speaker variation can be modeled. Speaker adaptation is achieved only by the constraint that the position in the speaker space is constant over each utterance. Therefore, no separate adaptation session is needed and the adaptation is present from the first utterance. Consequently, for a user there is no noticeable difference between this system and a speaker-independent system. The speaker model and the phonetic classification are implemented in the ANN part of a hybrid ANN/HMM system. In experiments with a pilot system, word accuracy is improved for utterances longer than three words and utterance level results are improved for utterances of all lengths.

1. INTRODUCTION

It is well known that one of the fundamental problems of automatic speech recognition (ASR), the large variability in the acoustic realization, can be reduced by adapting the speech recognition system to the user [4]. When the amount of calibration data from a speaker is large enough, all parameters of the ASR system can be re-estimated using the calibration data only, yielding a speaker-dependent (SD) system. In many cases however, it is not realistic to collect enough speaker specific calibration data for a complete re-estimation. A popular and theoretically appealing solution to this problem is to use Maximum a Posteriori (MAP) estimates of the parameters. This is an optimal weighting of the original, speaker-independent (SI) system's parameters and the parameters estimated from the calibration data [9]. A consequence is that the system continuously changes from an SI system when the amount of calibration data is small, to an SD system when the amount of calibration data is large. A nice property of the MAP method is that a statistical model of the speaker variation is defined by the prior density parameters, e.g., the mean and variance of each parameter (where the mean is the SI estimate of the parameter). A problem is that, for several reasons the parameters of the SI system define a poor model of the speaker variation. One reason is that the optimal dimension of the speaker space, modeling the speaker variation, is likely to be much smaller than the number of parameters in the SI system. There are at least two strong arguments for this. First, it is very unlikely that the speech code itself is less complex than individual speaker's variations of it. A consequence would be that learning to understand a new speaker would be as complex as learning a new language. Second, the amount of calibration data is less than the amount of training data used to estimate the SI parameters, and therefore it is only possible to robustly estimate fewer parameters. In response to this problem, parameter tying and linear regression techniques have been used where the large number of parameters are mapped to a smaller space by a simple transformation [5,9]. This leads to a second problem: there exists no simple transformation from the parameters of the SI system to an efficient model of the speaker variation. As an example, vocal tract length affects the acoustic realization of all voiced sounds in a very systematic fashion: the center frequencies of all formants are scaled by a speaker dependent factor. However, the mapping from formants to the features of the ASR system, e.g., mel cepstrum coefficients, is non-linear, and therefore there is no linear transformation from vocal tract length to the parameters of the SI system.

A different approach is to define a set of independent speaker parameters that models the speaker variation [7,8]. The speaker variation modeling is formally de-coupled from the rest of the ASR system and the parameters of this *speaker model* can be chosen to efficiently model the speaker variation. Efficient, in this case, means that a large part of the systematic variation can be modeled by a small number of parameters. To utilize the speaker model, the other components of the system are *speaker-sensitive*, i.e., their behavior is conditioned by the speaker parameters. Of course, the mapping from speaker parameters to variation in acoustic realization is still very complex. However, the speaker-sensitive behavior, of for example the acoustic pattern-matching component of the ASR system, can be trained off-line using large amount of data from many different speakers in a training database. Therefore, parameters of a complex mapping can be estimated. This framework is useful for speaker adaptation in general but the advantage of an efficient model of the speaker variation is greater when the amount of calibration data from a novel speaker is small. In particular, we believe that explicit speaker variation modeling is advantageous for unsupervised, instantaneous adaptation. In the next sections we describe a pilot ASR system with an explicit speaker model and present some recognition results for the Waxholm database [1].

2. THE BASELINE ASR SYSTEM

The baseline ASR system is a hybrid ANN/HMM system where the output probabilities of the HMM states are estimated by output activities of an ANN[2].

2.1. Lexicon and Grammar

The Markov model of the hybrid system uses the lexicon of the Waxholm dialogue project, which is a medium size (about 900 words) task-dependent lexicon with multiple pronunciations for many of the more common words [1]. A class-bigram grammar with perplexity 28 is used.

2.2. Feature Extraction

Mel cepstrum coefficients are extracted from a the short-time FFT spectrum of the speech signal every 10 ms. The speech is sampled at 16kHz and the short-time spectrum is computed with a Hamming window of 25 ms. The first 12 cepstrum coefficients, their first and second time-derivative and the log energy are the input features to the ANN of the hybrid system.

2.3. Phonetic Classification

An ANN with both recurrent connections and simple time-delay connections is used for the phonetic output probability estimation [2,7]. The input features are connected to a hidden layer of 40 units with a time-delay window of +/- 2 frames. Additionally, the hidden units are fully intra-connected with recurrent connections delayed one and two frames. Finally, the 45 phoneme output units are connected to the hidden units with a time-delay window of +/- 1 frame. The network was trained using the back-propagation through time algorithm [6] on 1418 training utterances of the Waxholm database. It has been shown that the activation of the output units estimates the class probabilities, $P(c_i \mid o)$, where c_i is phoneme *i* and *o* is the acoustic observation [3]. Bayes' rule gives:

$$P(o \mid c_i) = \frac{P(c_i \mid o)}{P(c_i)} P(o),$$

where $P(o)$, is constant for all competing hypotheses and $P(c_i)$ is estimated off-line from the training database. Therefore we can replace the output probabilities in the HMM with: $a_i / P(c_i)$, where a_i is the activation of the output unit corresponding to phoneme *i*.

3. SPEAKER VARIATION MODELING

The speaker variation modeling is incorporated in the ANN by adding *speaker space units* whose activities are identified with corresponding speaker parameters. Further, special purpose *speaker units* are introduced. One speaker unit per speaker in the training database is added, and the activity of a speaker unit is defined to be 1.0 when the corresponding speaker is the current speaker and 0.0 otherwise. The speaker units are connected to the speaker space units who are in turn connected to the hidden units. The topology of the speaker sensitive ANN is shown in Figure 1. The structure of this ANN has several appealing properties. The connections from and to the speaker space units are trained by back-propagation through time in the same optimization as all other connections of the network. The activation of a speaker space unit is a function of only the connection strength from the current speaker unit. Therefore, after training, the positions in the speaker space of all training speakers can be determined by inspecting these connections. Figure 2 shows the speaker space of the speaker sensitive ANN used in this study with two speaker space units. In [7] we analyze this automatically generated speaker space in detail. Here we just note that male and female speakers are effectively separated in the space.

The introduction of a speaker variation model in the ANN had an unanticipated positive effect on the network performance. If the units of the speaker model are removed and the biases of the hidden units are re-trained, the resulting SI ANN performs better than a SI ANN trained directly. Apparently, the ANN with a

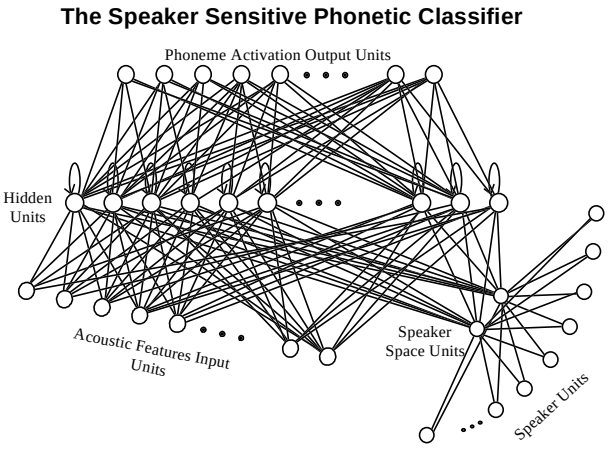

Figure 1: .The topology of the speaker sensitive ANN. See the main text for details.

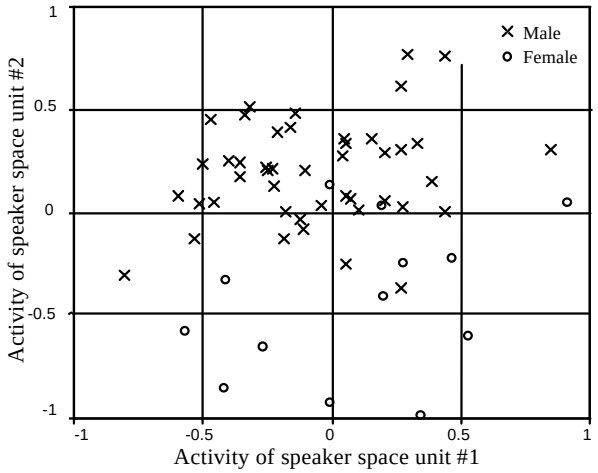

Figure 2: The automatically generated speaker space. Scatter-plot of the activation of the two speaker space units for all speakers in the training data after training.

speaker model learns better internal representations. This scheme was used to train the baseline SI ANN of this study but not in [7] which partly explains the greater difference between speaker sensitive and SI classification reported there.

4. INSTANTANEOUOS UNSUPERVISED SPEAKER ADAPTATION

The speaker-sensitive ANN can be used for different kinds of speaker adaptation. In this study instantaneous adaptation is investigated, i.e., no calibration data is used and the effect of adaptation is achieved by implementing the condition that the speaker, and thus the speaker parameters, are constant over each utterance. The algorithm is restarted for each new utterance (but see also sec. 5) and consequently, for a user there is no noticeable difference between this system and a speaker-independent system.

In a probabilistic formulation, speaker-independent ASR is to find the word-string w that maximizes the joint probability with an acoustic observation o of an utterance, i.e.,

$$\arg\max_{w}\left\{p(o|w)P(w)\right\}, \tag{1}$$

Similarly, instantaneous speaker adapting ASR can be written:

$$\arg\max_{w}\left\{P(w)\int_{\Lambda} p(o|w,\lambda)p(\lambda)d\lambda\right\}, \tag{2}$$

where Λ is the speaker space and λ is the speaker parameters, i.e., we assume that the speaker parameters are constant and integrate over the speaker space. It is easy to see that this is a generalization of SI ASR and reduces to (1) if o is independent of λ. The integral in (2) must be approximated to be practically useful. First we partition the speaker space into a number of regions i and make the approximation that $P(o \mid w, \lambda)$ is constant in each region. We get:

$$\arg\max_{w}\frac{P(w)}{N}\sum_{i=1}^{N} p(o|w,\lambda_i)p(\lambda_i)d\lambda, \tag{3}$$

where λ_i is the center of region i. This approximation can be implemented by generating an N-best list for each λ_i and sum the probabilities of entries in the lists with identical word-strings. Another approximation is:

$$\arg\max_{w}\left\{\max_{i} p_i(o|w,\lambda)P(w)p(\lambda_i)\right\} \tag{4}$$

where we, in analogy with the Viterbi approximation, replace the sum with its greatest term. Preliminary experiments showed no significant performance difference between the two approximations and for simplicity, the second approximation is used in all experiments below.

5. ACCUMULATING UNSUPERVISED ADAPTATION

If it can be determined from the dialogue when a new user starts to use the system, it is unnecessary to restart the adaptation for each new utterance. The probability of speaker space regions can be carried over to subsequent utterances. This can be formalized as follows:

$$\arg\max_{w}\left\{\max_{j}\left\{p(o_i|w,\lambda_j)P(w)\right\}q_{i,j}\right\} \tag{5}$$

$$q_{i,j}=\begin{cases} p(\lambda_j) & \text{if} \quad i=0 \\ \max_{w}\left\{p(o_i|w,\lambda_j)P(w)q_{i-1,j}\right\} & \text{otherwise} \end{cases} \tag{6}$$

where $q_{i,j}$ are the accumulated probabilities for the regions of the speaker space after utterance i, and o_i is the acoustic observation of the ith utterance. The approximation in the computation of the integral over the speaker space is analogous to (4).

6. RESULTS

6.1. Evaluation Procedure

Ten speakers (300 utterances) of the Waxholm database excluded from training are used for evaluation on the word and utterance levels. 9.5% of the utterances has at least one word that is not covered by the lexicon. Only true words are counted in the statistics, i.e., pauses, breath segments etc. are not counted.

Viterbi decoding is used in the evaluation of the baseline SI system. The speaker adapting system solves the integral of (2) using the approximation of (4). The same Viterbi decoding as in the SI case is used, but N=5 different points in the speaker space are evaluated in parallel. Note that the computational effort to perform the speaker sensitive decoding is not N-fold that of the SI because the same beam can be used in the beam-pruning for all N regions of the speaker space. Therefore, the decoding in points of the speaker space with low probability are pruned more heavily than in more probable points.

6.2. Instantaneous Adaptation

Over all utterances, there is no difference in word recognition performance for the SI and speaker adapting system as can be seen in Table 1. However, an analysis of the performance for different utterance lengths shows that the speaker adapting system performs worse than the SI system only for short utterances of one, two and three words (Figure 3). At all other lengths the speaker adaptation improves the accuracy. The general pattern that adaptation is more powerful for longer utterances was expected and the break-even point between three and four words is quite low. The lack of overall improvement can be explained by the large proportion of short utterances in the database.

Utterance recognition is not degraded even for very short utterances (see Figure 4 and Table 1). The difference between word and utterance levels is not surprising as the adaptation is based on an additional utterance level constraint.

6.3. Accumulating Adaptation

The experiments with accumulating adaptation are similar to the instantaneous adaptation. The speaker space is partitioned into the same five points, but the adaptation is only restarted once for each test speaker. This is an easier problem than the instantaneous adaptation and Table 1 shows that both word accuracy and utterance recognition are improved.

7. CONCLUSIONS

A simple pilot continuous speech ASR system was used to evaluate how explicit speaker variation modeling can be utilized for speaker adaptation. Although a very small ANN was used, the speaker space dimension was low (2 parameters) and the sampling of the speaker space was coarse, the results show that fast adaptation is possible (positive effect after four spoken words). This is very promising, but further research is needed to answer how the method scales up with more complex systems and speaker spaces.

	Speaker Independent	*Instantaneous Adaptation*	*Accumulating Adaptation*
Word Accuracy	72.4 %	72.4 %	73.1 %
Correct Utterances	43.0 %	45.0 %	44.0 %

Table 1: Recognition results for the speaker independent baseline system and adaptation. Word accuracy includes substitution, insertion and deletion errors.

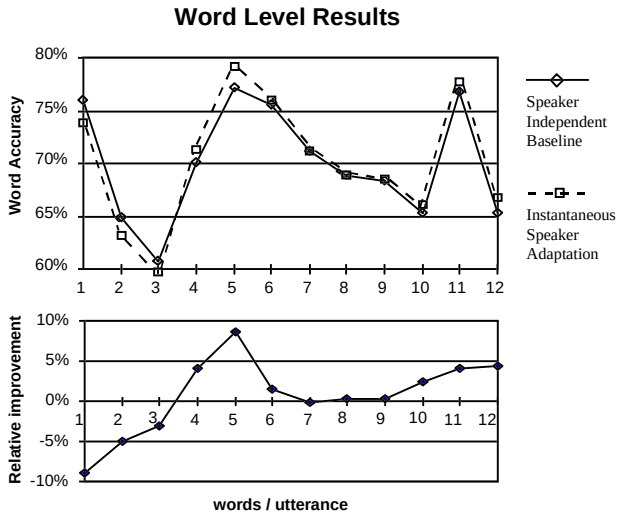

Figure 3: Top: Word accuracy as a function of utterance length. Bottom: Relative improvement from SI to speaker adaptation.

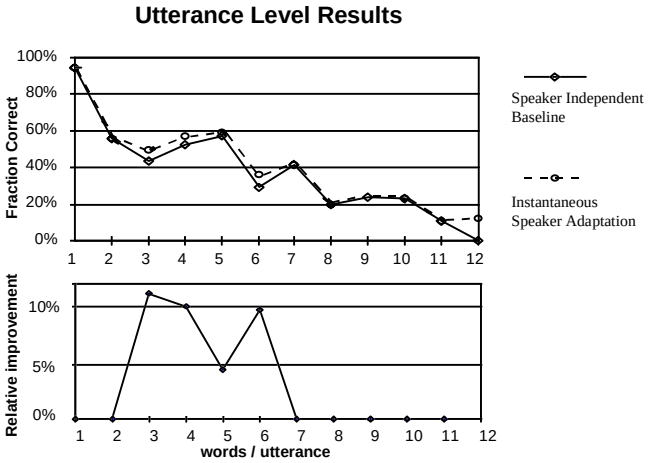

Figure 4: Top: Fraction utterances correct as a function of utterance length. Bottom: Relative improvement from SI to speaker adaptation.

REFERENCES

1. Bertenstam, J., Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa-Leitao, A. & Ström, N. "The Waxholm Application Database," *Proc. EUROSPEECH '95* pp.833-836, 1995.

2. Bourlard, H. & Morgan, N., *Continuous Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, 1993.

3. Gish, H. "A Probabilistic Approach to the Understanding and Training of Neural Network Classifiers," *Proc. ICASSP '90*, pp. 1361-1364, 1990.

4. Huang, X.D. & Lee, K.F. "On Speaker-Independent, Speaker-Dependent and Speaker-Adaptive Speech Recognition," *Proc. ICASSP '91.*, pp. 877-880, 1991.

5. Leggetter, C.J. & Woodlland, P.C. "Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression," *Proc. ICSLP '94*, pp 451-454, 1994.

6. Pealmutter, B.A. "Dynamic Recurrent Neural Networks," *TR CMU-CS-88-191*, CMU Comp. Sc. Dept., 1990.

7. Ström, N. "A Speaker Sensitive Artificial Neural Network Architechture for Speaker Adaptation," *ATR TR-IT-0116*, ATR, Japan, 1995.

8. Ström, N. "Experiments with a New Algorithm for Fast Speaker Adaptation," *Proc. ICSLP '94*, pp. 459-462, 1994.

9. Zavaliagkos, G., Schwarz, R., McDonogh, J. & Makhoul J. "Adaptation Algorithms for Large Scale HMM Recognizers," *Proc EUROSPEECH '95*, pp 1131-1134, 1995.