A New Paradigm for Speaker-Independent Training and Speaker Adaptation

Francis Kubala and Richard Schwartz

BBN Systems and Technologies Corporation
Cambridge, MA 02138

Abstract

This paper reports on two contributions to large vocabulary continuous speech recognition. First, we present a new paradigm for speaker-independent (SI) training of hidden Markov models (HMM), which uses a large amount of speech from a few speakers instead of the traditional practice of using a little speech from many speakers. In addition, combination of the training speakers is done by averaging the statistics of independently trained models rather than the usual pooling of all the speech data from many speakers prior to training. With only 12 training speakers for SI recognition, we achieved a 7.5% word error rate on a standard grammar and test set from the DARPA Resource Management corpus. This performance is comparable to our best condition for this test suite, using 109 training speakers.

Second, we show a significant improvement for speaker adaptation (SA) using the new SI corpus and a small amount of speech from the new (target) speaker. A probabilistic spectral mapping is estimated independently for each training (reference) speaker and the target speaker. Each reference model is transformed to the space of the target speaker and combined by averaging. Using only 40 utterances from the target speaker for adaptation, the error rate dropped to 4.1% — a 45% reduction in error compared to the SI result.

1. Introduction

One important scenario for the use of spoken language systems (SLS) by new speakers is to start with a SI corpus or model and have the system adapt as the new users interact with the system. Once the interaction has begun, the system has the opportunity to collect speaker-dependent data of known orthographic transcription from the target speaker. After a small sample of speech has been collected, the system should be able to adapt so as to significantly increase performance compared to the original SI model. The success of this scenario depends on the adaptation being powerful enough to generalize from a small sample of speaker-specific speech in which most of the phonetic contexts of the language are not observed. Furthermore, it depends on having a SI speech corpus which is amenable to speaker adaptation.

It is a widely held belief that speech used for training SI models must be collected from many speakers. It is also commonly accepted that collecting only a small sample of speech from each training speaker is a reasonable compromise to make in the effort to collect as many speakers as possible. While this compromise may be reasonable for SI recognition, several efforts to use such a corpus as a basis for speaker adaptation have failed to make significant improvements.

Recently, we have discovered that adequate SI performance can be achieved with far less speaker coverage than conventionally thought necessary, but with much better sampling of each training speaker's speech. Specifically, we show that it is possible to achieve near state-of-the-art SI performance on a 1000-word continuous speech recognition task using only 12 training speakers. Furthermore, we will show that it is possible and advantageous to create the SI model from a set of independently trained speaker-dependent (SD) models, without retraining on the entire pooled dataset at one time. Most importantly, we show that such a SI corpus is an effective basis for speaker adaptation. By combining the adapted models of 11 reference speakers, we were able to reduce the error rate by 45% compared to the SI performance. This method succeeds because we are able to apply a robust probabilistic speaker–transformation to well-trained and highly discriminating SD training models.

In section 2, we describe the new SI training paradigm and present comparative results for SI recognition using only 12 training speakers. In section 3, we describe three previous attempts to adapt from a corpus of many training speakers. Then we describe our approach for adapting to new speakers from the 12 speaker SI corpus and discuss experimental results.

2. Speaker-Independent Training

109 Speaker SI Training

For several years, the DARPA Resource Management continuous speech corpus has provided a testbed for SI recognition. 109 speakers are designated as training speakers and are each represented by a sample of 40 utterances. Typically, the data from all the speakers is pooled at the outset, as if it all came from one speaker. Although the training data originates from many diverse sources, the forward-backward (Baum–Welch) training procedure is robust enough to do a reasonable job of modeling the pooled data. When used with a standard word-pair grammar of perplexity 60, state-of-the-art SI recognition performance for this corpus is 6–7% word error rate.

This performance is 3 times worse than our current SD performance using 600 training utterances. Also, the sentence error rate at this level of performance is greater than 30% — a level of error that we assume is far too high for the acoustic component of a spoken language system. Furthermore, this performance has been achieved with an artificial and non-robust grammar of modest perplexity which will not work within an SLS context. Combining the need for higher absolute performance with the need to use less powerful grammars indicates that the current SI error rate may need to be reduced by a factor of at least 4 to be acceptable for SLS applications.

12 Speaker SI Training

Since we planned to perform adaptation from 12 reference speakers, we needed to run a SI control condition by using the data in the usual pooled fashion. We ran a comparative test using data from only the 12 speakers from the SD segment of the DARPA database. The training for each speaker consisted of 600 training utterances. Seven of the speakers are male.

We did have some indication that pooling the data of even a few speakers could make large improvements from an experiment conducted at IBM and described in [5]. However, 12 speakers could hardly be expected to contain an example of all speaker types in the general population (including both genders), so we could anticipate the need for some kind of smoothing before we began. Our usual technique for smoothing across the bins of the discrete densities, triphone cooccurrence smoothing [7], has proven to be an effective method for dealing with the widely varying amounts of training data for the detailed context models in the system. When used in a SD training scenario, it has allowed us to observe a performance gain for explicitly modeling several thousand triphones which were observed only once or twice in the training.

However, the cooccurrence smoothing is not appropriate for models derived from the pooled data of many speakers. Spectra from different speakers will cooccur much more randomly than spectra from a single speaker. This will yield poorer estimates of the smoothing matrices. As such, triphone cooccurrence smoothing is a *speaker-specific* modeling technique. If the data is pooled prior to training, we cannot effectively apply our best smoothing to the model.

This realization has led us to examine the practice of pooling the data in the first place. A straightforward alternative to pooling the data is to keep the speakers separated until the speaker-specific operations of training and smoothing have been completed and then combine the multiple SD models. To allow the model combination to be done by averaging the model statistics, we constructed a SI codebook which was used in common for all speakers.

Results of SI Experiments

Results for several SI experiment are shown in table 1. All results are from first runs of the designated Feb. '89 SI test set on the given system configuration. This test set consists of 10 speakers (4 females) with 30 utterances each. All runs used the standard word-pair grammar of perplexity 60. System parameters were fixed before running any of the conditions in this experiment. The limited development testing which we did perform was done only on the June '88 SD/SI test set using only the 109 speaker SI model.

For each condition we show the number of training speakers, and the manner in which the models were trained and smoothed. The training was done either on pooled data (*joint* training) or on individual speakers' data (*indep* training). The smoothing was either not done, or was applied to either the jointly or independently trained model. For each condition, the word error rate, which includes insertion errors, and sentence error rate are given.

#Spkrs	Training	Smoothing	Word Err	Sent Err
109	joint	none	7.1	36
109	joint	joint	6.5	34
12	joint	none	9.0	42
12	joint	joint	8.5	41
12	joint	indep	7.8	37
12	indep	indep	7.5	37

Table 1. Comparison of SI training scenarios on the Feb. '89 test set with word-pair grammar.

The 109 speaker conditions were run to calibrate the BYBLOS system with published results for the same test set. We observe a small improvement, from 7.1% to 6.5% word error, for using smoothing on the jointly trained model. The 6.5% error rate is comparable to the best performance on record (6.1%) for this test set which was achieved by Lee as noted in [4]. Furthermore, the sentence error rates are identical. Lee's system used a corrective training and reinforcement procedure to increase the discrimination ability of the model for confusable words. No corrective training was used for the BYBLOS results given in table 1.

The system configuration for the 109 condition was identical to that which we use for SD recognition except for one difference. One new system parameter was added to decrease the lambda factors used for combining the context-dependent models into interpolated triphones [6] by a factor of eight to account for the larger corpus.

Next we repeated the same conditions for the 12 speaker SI model. Simply pooling the 12 speakers without smoothing does not perform as well as the 109 speaker model. And once again, smoothing the jointly trained model has a rather weak effect on performance. However, we were surprised that the 12 speaker model should have only 25% more error than the 109 speaker model.

The final two results show the effect of independently smoothing the 12 speaker model after either joint or independent training. To independently smooth the jointly trained model, we first trained on the pooled data as usual. Then a

SD model was made, for each training speaker, by running the forward-backward algorithm on the combined SI model but on data from only one speaker in turn. This allowed us to generate a set of SD models for smoothing, which shared a common alignment. The smoothed models were then recombined by averaging the model statistics.

The approach used on the final result is the most straightforward — we train multiple independent SD models allowing each to align optimally for the specific speaker, smooth each model to model spectral variation within each speaker, and then combine the models by averaging corresponding probabilities in the models.

As is evident from the table, both of the final methods improve due to the increased effectiveness of the smoothing when it is applied to a speaker-specific model. In a final surprise, we find that constraining all the speakers to a common alignment does not help. Further, the word error rate of this simple model is only 15% worse than our best performance with the 109 speaker model and the sentence error rates are statistically indistinguishable.

Some caution is required in comparing results of the 12 and 109 speaker models due to two, possibly important differences. The total amount of training speech used is different as is the number of different sentence texts contained in the training script. The 109 speaker model is trained on a total of 4360 utterances drawn from 2800 sentence texts. The 12 speaker model is trained on 7200 utterances drawn from only 600 sentence texts. While the additional speech may benefit the 12 speaker condition, the greater richness of the sentence texts may help the 109 speaker model. The effect of the additional sentence texts can be seen in the different numbers of triphone contexts observed in the two training scripts: 5000 triphones for 600 sentences vs. 7000 for the 2800–sentence script.

Discussion of SI Results

We have observed that the forward-backward algorithm freely re-defines some of the phonemes to model peculiarities of a given speaker. If we constrain all speakers to a common alignment, the training procedure must make a compromise between these speaker-specific adjustments. Both forward-backward and triphone coocurrence smoothing are arguably speaker-specific procedures — they work best when the training distributions are generated by a single source. Some compromise must be made for SI recognition, where the training is not homogeneous and the test distribution is, by definition, different than the training. It appears, from these results, that the least damaging compromise may be to delay pooling of the data/models until the last possible stage in the processing.

Such a simple SI paradigm has several attractive attributes. It makes the data collection effort easier. It is trivial to add new training speakers to the SI model; no retraining is required. Therefore the system can easily make use of any speakers who have already committed to giving enough speech to train a high-performance SD model. There is a large payoff for being one of the training speakers in this scenario — highly accurate SD performance. In contrast, there is no benefit for being a training speaker for the 109 speaker model. Finally, by delaying the stage at which the data or model parameters are pooled, new opportunities arise to use speaker-specific modeling approaches such as the multiple-reference adaptation procedure described in the next section.

3. Speaker Adaptation

Adaptation from 109 Speakers

As mentioned above, previous attempts to use large population SI corpora for speaker adaptation have met with little success. In [3], Lee tried to cluster over 100 training speakers into a small number of groups which were then trained independently. In recognition, the test speaker was first classified into one of the speaker groups, based on 1 known utterance, and decoded with the appropriate model. This approach failed to improve over the SI performance since it reduced the amount of training data available to each speaker-group-specific model. In another attempt, Lee devised an interpolated re-estimation procedure which combined the SI model with 4 other models derived from a small sample of known speech from the target speaker. Interpolation weights for the 5 models were computed from a deleted sample of the training data. The reduction in word error rate was less than 10%, however, when 30 utterances from the target speaker were used. The gain was small for this approach because only a small amount of new information, robustly estimated in the 4 speaker-specific models, was added to an already robust SI model.

We have also attempted to use the same SI corpus of over 100 speakers for speaker adaptation as reported in [2]. In this work, we estimated a deterministic transformation on the speech parameters of each of the training speakers which projected them onto the feature space of a single *prototypical* training speaker. We then trained on all of the transformed speech as if it came from a single speaker. The target speaker was similarly projected onto the prototypical speaker and recognition proceeded using the prototypical model. This procedure reduced the word error rate by 10% compared to the SI result; a minor improvement for a significant increase in the complexity of the scenario. We believe that this method did no better because the feature transformation was not powerful enough to superimpose a pair of speakers without significant loss of information. This resulted in a prototypical model whose densities were not significantly sharper than the comparable SI model made from the original data.

Adaptation from 12 Speakers

Our experience with the 109 corpus led us to rethink our approach to speaker adaptation from multiple reference speakers.

We already have a powerful speaker adaptation procedure which effectively transforms a single well-trained SD reference model into an adapted model of the target speaker [1]. The transformation is estimated from a small amount of adaptation data (40 utterances) given by the target speaker. The approach is powerful for two reasons: first, the estimate of the probabilistic spectral mapping between two speakers is robust and generalizes well to phonetic contexts not observed in the adaptation speech, and second, the transformation can be applied to the well-estimated, discriminating densities of the SD reference model without undue loss of detail.

A natural extension of this approach to multiple references would be to combine the parameters of several SD models after they had been independently adapted to the same target speaker. We can assume from our 12 speaker SI experiments that the transformation will perform better if estimated independently between each speaker-pair in turn rather than from a pooled dataset, since the transformation is a speaker-pair-specific operation. We also know that we can successfully combine the multiple adapted models by averaging the model statistics.

Results of Adaptation Experiments

Table 2 shows results for development tests on the June '88 SD/SI test set and word-pair grammar. The test set consists of 12 speakers (7 males) and 25 utterances each.

#Refs	Training	Word Err	Sent Err
1	30 min	6.2	31
1	2 hr	5.6	30
11	5.5 hr	4.1	23

Table 2. Comparison of speaker-adapatation results on the June '88 test set with word-pair grammar.

Adapting from a single male reference speaker trained on 30 minutes of speech (600 utterances) gives a word error rate of 6.2%. The reference speaker in this case is, LPN, from the designated RM2 database.

In the second row, a small improvement is realized for increasing the reference speaker training to 2 hours (2400 utterances). We intend to make this comparison more reliable by using the three other speakers in the RM2 database as references.

The third condition shows the result of combining models from 11 reference speakers after adapting them to the 12th speaker and jackknifing over all the reference speakers. The result is a significant improvement in both word and sentence error rates over the single reference performance.

Discussion of Adaptation Results

Speaker adaptation from a single reference speaker continues to be an economical solution for systems which are forced to retrain due to changes in channel, environmental conditions, or task domain. With only 40 utterances from the system users and 600 training utterances from the reference speaker, a speaker-adaptive system can be rapidly re-configured and deliver performance equal to the best current SI performance trained on 4000 utterances.

We can also make a comparison between the multi-reference adapted result, tested on the June '88 SD/SI test set, and the 12 speaker SI result tested on the Feb. '89 SI test set, since roughly the same population of training speakers are used (except for the held-out one). The two test sets give the same performance when tested using the 109 speaker SI model. Comparing to the 12 speaker SI model, the 11 reference adapted model has reduced the word error by 45%.

We are encouraged by this large improvement for a straightforward application of our basic speaker adaptation algorithm to multiple references. Individual speaker performance ranged from 0.6% to 7.7% error indicating that the multiple-reference model was very effective at eliminating the poorest outliers. Two speakers performed equal to or better than their SD models trained on 600 utterances.

We intend to continue investigating the potential of speaker adaptation from multiple references. If we can continue to improve our adaptation algorithm, and understand what constitutes good reference speakers, it may be possible to bring our speaker-adaptive performance very close to our SD performance.

Conclusions

We have shown that it is possible to achieve near current state-of-the-art SI performance with a model trained from only 12 speakers. This result is possible due to two important changes to the usual SI training paradigm — a large amount of speech is available from each training speaker and the data is not pooled before training.

Having a large sample of data from each training speaker and keeping it separate allows us to train detailed, highly discriminating, densities in a SD model and make the most effective use of *speaker-specific* modeling techniques such as triphone cooccurrence smoothing and probabilistic spectral transformation.

Furthermore, the new paradigm eases the burden of data collection for SI recognition and allows new training speakers to be added to the SI model with ease.

Most importantly, the new SI corpus lends itself well to speaker adaptation. By combining multiple reference speaker models which have been independently transformed to the target speaker, we have cut the SI word error rate from 7.5% to 4.1% using only 40 utterances of adaptation speech.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Office of Naval Research under Contract No. N00014-85-C-0279.

References

[1] Feng, M., F. Kubala, R. Schwartz, J. Makhoul, "Improved Speaker Adaptation Using Text Dependent Spectral Mappings", *IEEE ICASSP-88*, paper S3.9.

[2] Kubala, F., R. Schwartz, C. Barry, "Speaker Adaptation from a Speaker-Independent Training Corpus", *IEEE ICASSP-90*, Apr. 1990, paper S3.3.

[3] Lee, K., "Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System", PHD dissertation, Carnegie-Mellon University, Apr. 1988, CMU-CS-88-148

[4] Lee, K., H. Won, M. Hwang, "Recent Progress in the Sphinx Speech Recognition System", *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Inc., Feb. 1989, pp. 125–130.

[5] Rtischev, D., "Speaker Adaptation in a Large-Vocabulary Speech Recognition System", Masters thesis, Massachusetts Institute of Technology, Jan. 1989.

[6] Schwartz, R., Y. Chow, O. Kimball, S. Roucos, M. Krasner, J. Makhoul, "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech"; *IEEE ICASSP-85*, Mar. 1985, paper 31.3

[7] Schwartz, R., O. Kimball, F. Kubala, M. Feng, Y. Chow, C. Barry, J. Makhoul, "Robust Smoothing Methods for Discrete Hidden Markov Models", *IEEE ICASSP-89*, May 1989, paper S10b.9.