SPEAKER-TRAINED RECOGNITION USING ALLOPHONIC ENROLLMENT MODELS

V. Vanhoucke[†⋆], *M.M. Hochberg*[⋆], *C.J. Leggetter*[⋆]

[†] Department of Electrical Engineering, Stanford University, Stanford, CA
[⋆] Nuance Communications, 1380 Willow Road, Menlo Park, CA

ABSTRACT

We introduce a method for performing speaker-trained recognition based on context-dependent allophone models from a large-vocabulary, speaker-independent recognition system. In this approach, a set of speaker-enrollment templates is selected from the context-dependent allophone models. These templates are used to build representations of the speaker-enrolled utterances. The advantages of this approach include improved performance and portability of the enrollments across different acoustic models.

We describe the approach used to select the enrollment templates and how to apply them to speaker-trained recognition. The approach has been evaluated on an over-the-telephone, voice-activated dialing task and shows significant performance improvements over techniques based on context-independent phone models or general acoustic model templates. In addition, the portability of enrollments from one model set to another is shown to result in almost no performance degradation.

1. INTRODUCTION

In general, speech recognition systems do not rely on acoustic information to derive their language models. However, some applications require the ability to personalize the vocabulary interactively, without any medium other than the audio channel. In these applications, the user needs to be able to update the language model by interacting with the system by voice. As a consequence, the recognition engine needs to be able to update its models based solely on the acoustics collected during the interaction. A typical example would be a voice-activated dialing application in which users would be able to maintain a personal list of names. To add a name to the list, the user says the name one or more times and the recognition engine builds a model of the spoken utterance. This utterance model can then be used by the recognition system in a later interaction.

1.1. The Speaker-Trained Recognition (STR) Process

In order to add a new word or phrase to the system, the user is queried to speak it one or several times. The *enrollment* mechanism extracts the information needed by the recognizer and stores it in its database. When the word is subsequently used, during the *recognition* phase, the system retrieves the information from the database and combines it with the recognition models.

Since the enrollment operation modifies the recognition models online, several checks are performed to ensure that the recognition accuracy is not hurt by adding a new word. First, the system checks that the word, or any similar sounding one, is not already present in the language model. This operation is referred to as *clash testing*. In a second step, the validity of the acoustics of the word itself is checked. Typically the user is required to repeat the word at least twice. The system makes sure that the two utterances are consistent with each other. *Consistency testing* ensures that the portion of speech captured by the system is indeed the one intended to be enrolled.

1.2. Design Constraints

In the design of a STR system, several factors come into play beyond the typical accuracy issues.

From a user interface perspective, the enrollment operation must be fast and easy to perform. This means that the false rejection rate of both clash and consistency testing has to be minimized, while ensuring that the false accept rate is low enough to maintain the usefulness of the tests.

Because of the distinctive origin of words enrolled using STR, the accuracy of the system on these words considered in isolation might not be consistent with the performance when used in conjunction with speaker-independent grammars. Maintaining a high level of accuracy in speaker-independent contexts is critical to applications such as voice-activated dialing.

In addition, it is of practical importance for the compatibility of enrolled words to be maintained when the underlying acoustic models are changed or upgraded. This ensures that improvements obtained through algorithmic changes, retraining or adaptation of the models will not be detrimental to speaker trained words, and that the interoperability with speaker-independent tasks is consistently maintained.

2. BASELINE SYSTEMS

Typical speaker-trained recognition systems fall into two categories: *phonetic systems*, which use phone-like bases as enrollment models ([1], [2], [3]), and *acoustic systems*, which use low-level representations of the acoustics (e.g. DTW or template matching systems; see also [4]).

2.1. Baseline Phonetic System

Phonetic systems use phonetically labeled models as templates for enrollment. In practice, these systems generally rely on broad speaker-independent models, which can undermine their accuracy when enrollments are combined with large vocabulary speaker-independent grammars.

The baseline phonetic system considered in this paper uses a set of monophone hidden Markov models to represent speech. The enrollment process learns the sequence of monophone models corresponding to the enrolled utterance. Two consistent pronunciations of the words are required, and the phoneme sequences generated are used as alternative pronunciations of the enrolled word in a similar way as the one proposed in [3]. At recognition time, the monophone enrollment models are evaluated to determine the spoken utterance. An advantage of the monophone approach is that the number of parameters involved is minimal, leading to a very inexpensive enrollment process.

2.2. Baseline Acoustic System

Acoustic systems take full advantage of speaker-dependence, and model the acoustics of the training data at a fine level with no explicit phonetic constraints. These are in general very accurate, and can be made robust by training the templates on speaker-independent data. The major drawback of these systems is portability: since the templates used are explicitly tied to the underlying acoustic models, robustness to modifications or adaptation of the underlying modelset is poor.

The baseline acoustic system evaluated in this paper is built on top of the phonetic baseline. The underlying acoustic model considered uses genones [5] as acoustic model clusters. A single enrollment model is derived from each cluster by averaging the mixture weights of all the allophones pointing to this particular genone. This method ensures a proper coverage of the complete acoustic space, while limiting the total number of enrollment models to a reasonable level. These models are then added to the enrollment grammar loop in parallel with the monophones.

3. PROPOSED APPROACH

3.1. Allophonic Enrollment Models (AEM)

A natural way to combine the advantages of both acoustic and phonetic systems is to consider using speaker-independent, context-dependent allophonic models for enrollment. Allophones have a labeling that depends only on the phone set, potentially providing some level of independence with respect to the underlying acoustic models. On the other hand, large vocabulary speech recognition systems typically use several thousand models, providing a fine segmentation of the acoustic space.

The proposed system uses an unconstrained loop of allophones for enrollment. An additional model transition penalty is applied during enrollment to control the length of the allophone sequence. Because the objective is to obtain a transcription of the spoken utterance at a potentially finer level than the phonetic transcription, none of the context dependencies between allophones are enforced. This policy lets each state of each model align to whichever segment of the acoustic string is the most likely, regardless of the neighboring allophones, and provides a richer set of templates to match against at any point in time. This reduces dramatically the complexity of the enrollment grammar and significantly eases the process of selecting the enrollment models.

3.2. Model Selection

The grammar used to enroll speaker-trained items is an unconstrained AEM loop. As a consequence, the enrollment speed is inversely proportional to the number of models used. To make this approach practical, only a subset of the available pool of allophones will be considered.

A natural way of selecting the appropriate AEM set is to consider a loop containing all the available models, enroll a large amount of data with it, and select the models with highest unigram probabilities. Figure 1 shows the frequency/rank distribution of the models for a typical speaker-independent modelset (dotted line is an exponential fit). Note that the model distribution approximately follows Zipf's law, with a smaller tail than the law would predict. This indicates that very good coverage of the models actually used during enrollment can be achieved while selecting relatively few of them.

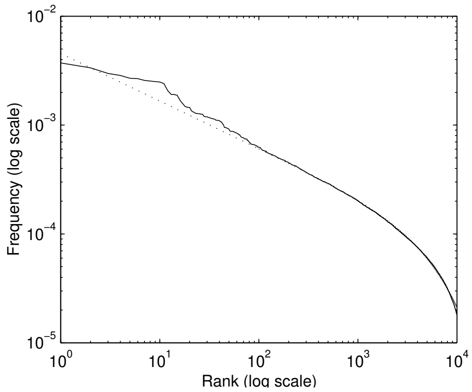

Fig. 1. Frequency/rank distribution of the AEM

This result can be assumed to apply to any type of mixture model. In the context of genonic mixture models, the genone coverage can be expected to be a good statistic of the level of coverage of the acoustic space. Figure 2 shows that although the percentage of the training data covered by selecting a subset of the AEM grows slowly with the rank of the last AEM selected, very good coverage of the genone and phonetic spaces can be achieved with very few AEM models.

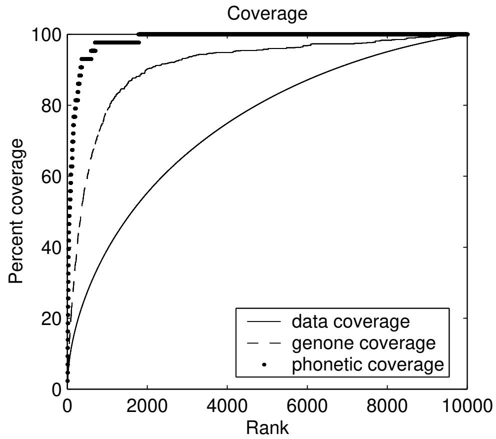

Fig. 2. Coverage Statistics

There are some practical issues associated with this selection process. Since a significant portion of the probability is contained in the low-ranking models, how this probability mass redistributes across the top-ranking models after selection might significantly alter their ordering. Ideally, selection should be carried out in successive steps, reevaluating the distribution of the models on the data after each resampling.

An alternative scheme is to consider the allophone statistics collected during the training of the speaker-independent acoustic models. These statistics are a by-product of the training process and there is no overhead associated with collecting this information. They are only consistent across allophones with the same context span. It is reasonable to consider only the longest context span allophones (typically triphones) and select the ones with the highest prior probability. These models have been trained on the largest amount of data and are expected to be best at segmenting the acoustic space.

This second selection method is also a reasonable predictor of the previous one. On an American English modelset, the Spearman rank correlation between the two selection schemes is about 0.89, which means that the ranking of the models obtained during training is a good predictor of the ranking obtained by enrolling data. The two selection methods have also shown experimentally to perform similarly.

4. EXPERIMENTAL SETUP

The task used to evaluate the AEM STR approach simulates a voice-activated dialing application. The data consists of 66 American speakers enrolling over the telephone channel 50 names each. Each enrollment corresponds to two consistent repetitions of the name. Recognition accuracy is evaluated on 50 more utterances per speaker. Clash testing is disabled during accuracy experiments in order to keep the number of utterances constant.

Clash testing is implemented using an approximate rescoring of the acoustics of the word being enrolled against the pronunciations of the potential clashes. Words above a threshold on the likelihood distance between the current enrollment and the other words are considered clashes. Clash testing performance is evaluated by attempting to enroll the the same name several times under different labels (False Accepts), and by counting the number of enrollments that were rejected although they were very dissimilar acoustically (False Rejects).

Consistency testing is implemented using an approximate rescoring of the acoustics of the word being enrolled against the pronunciations of the words to be consistent with. Words below a threshold on the likelihood distance between the current enrollment and the other words are deemed inconsistent. Consistency testing performance is evaluated by trying to enroll under the same label utterances whose acoustics are very dissimilar (False Accepts), and by counting the number of enrollments that were rejected although they were distinct utterances of the same name (False Rejects).

Robustness to combination with speaker-independent grammars is evaluated by running the same recognition experiment, adding in parallel with the STR names a list of SI names of variable size.

Models	Error Rate
Monophones	2.9 %
1000 Genones	2.7 %
1000 AEM	2.1 %

Table 1. STR Accuracy

5. RESULTS

5.1. Accuracy

Table 1 compares the error rate of the various setups on the STR task without any competing speaker-independent entries. Performance of the AEM system is 22% better than the system based on genonic templates, and 27% better than the monophone system.

Experiments combining speaker-independent and STR name lists demonstrate clearly the drawbacks of typical phonetic approaches (Figure 3). The performance of a monophone loop does not scale well with the number of names. While the size of the enrollment modelset does not seem to matter in isolation, for any practical application using STR in a speaker-independent context, the granularity of the acoustic models is critical. This is demonstrated by the genones and AEM performances as the name list size grows.

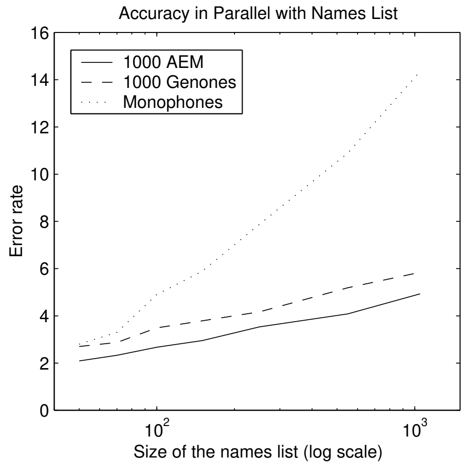

Fig. 3. Combining STR and SI Grammars

5.2. Clash and Consistency Testing

Performance of clash testing is greatly improved by using AEM for enrollment (Figure 4). At typical operating points (low false reject rates), the reduction in false accepts at a given false reject rate is about twofold over genones and monophones.

Performance of consistency testing (Figure 5) follows a similar trend. At low flase reject rates, the reduction in false accept rates compared to the baseline systems is even more dramatic (around 70%).

5.3. Robustness to changes in the underlying acoustic model

In this experiment, two different modelsets (denoted Model I and II in Table 2) were used to evaluate the portability of AEM across acoustic models. The two modelsets have very different model parameter complexities. The training data for both models overlap, but the training process for each of them was different enough to simulate portability across successive retrainings of models.

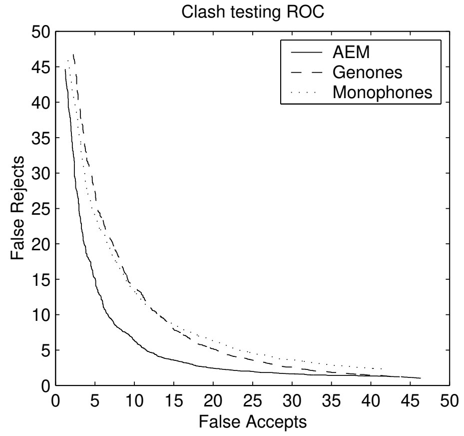

Fig. 4. Clash Testing ROC

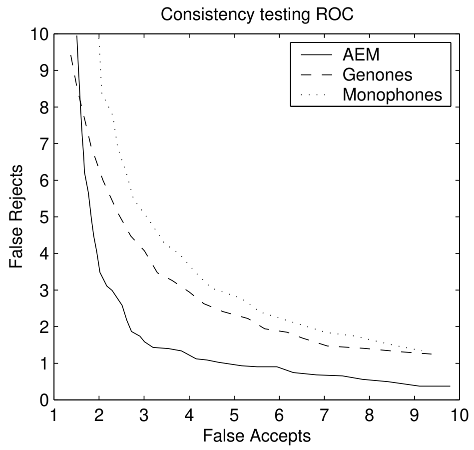

Fig. 5. Consistency Testing ROC

Table 2 shows the result of porting enrollments on one modelset to the other. The porting of enrolled words between modelsets, which was impossible with an acoustic enrollment method, can be performed at no statistically significant cost when AEM are used as templates.

Enrollment	Recognition	Error Rate
Model I	Model I	2.68%
Model I	Model II	2.71%

Table 2. Robustness to Modelset Change

6. CONCLUSION

We showed that the use of Allophonic Enrollment Models provides a robust, practical method to perform speaker-trained recognition. The main features of the technique are an excellent tradeoff between speed and accuracy, robustness to speaker-independent environments and portability across modelsets. In addition, this technique is very simple to implement on top of a speaker-independent speech recognition engine. It guarantees that improvements to the performance of the speaker-independent system, through algorithmic changes as well as improvements of the acoustic models, can be integrated into a system containing speaker-trained enrollments without degrading STR performance.

7. REFERENCES

[1] N. Jain, R. Cole, and E. Barnard, “Creating speaker-specific phonetic templates with a speaker-independent phonetic recognizer: Implications for voice dialing,” *Proceedings of ICASSP 96*, pp. 881–884, 1996.

[2] V. Fontaine and E. Bourlard, “Speaker-dependent speech recognition based on phone-like units models — application to voice dialing,” *IDIAP-RR 96-09*, 1996.

[3] T. Thomas B. T. Tan, Y. Gu, “Implementation and evaluation of a voice-activated dialling system,” *Proceedings of IVTTA 98*, pp. 83–86, 1998.

[4] V. Raman and V. Ramanujam, “Robustness issues and solutions in speech recognition based telephony services,” *Proceedings of ICASSP 97*, pp. 1523–1526, 1997.

[5] V. Digalakis, P. Monaco, and H. Murveit, “Genones: Generalized mixture tying in continuous hidden markov model-based speech recognizers,” *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 281–289, 1996.