A UNIFIED APPROACH OF INCORPORATING GENERAL FEATURES IN DECISION TREE BASED ACOUSTIC MODELING

Wolfgang Reichl and Wu Chou

Bell Laboratories, Lucent Technologies,
600 Mountain Ave.,
Murray Hill, NJ 07974, USA

ABSTRACT

In this paper, a unified maximum likelihood framework of incorporating phonetic and non-phonetic features in decision tree based acoustic modeling is proposed. Unlike phonetic features, non-phonetic features in this context are those features, which cannot be derived from the phoneme identities. Although non-phonetic features are used in speech recognition, they are often treated separately and based on various heuristics. In our approach, non-phonetic features are included as additional tags to the decision tree clustering. Moreover, the proposed tagged decision tree is based on the full training data, and therefore, it alleviates the problem of training data depletion in building specific feature dependent acoustic models. Experimental results indicate that up to 10% word error rate reduction can be achieved in a large vocabulary (Wall Street Journal) speech recognition task based on the proposed approach.

1. INTRODUCTION

In speech recognition, many non-phonetic features are used to improve the resolution of the acoustic model and to obtain high recognition performance. Examples of such features include gender, speaker or speaker group identity, speaking rate, channel and environment condition, ambient noise level, etc. However, these features are not phonetic features in the sense that they cannot be derived from the phoneme identities and it has been a problem of how to incorporate them consistently with phonetic features in high-resolution acoustic modeling. The common practice is to manually separate the data according to the specification of the non-phonetic features, such as gender, and retrain a model using only the data which posses these features. This approach has two major problems. First, it depletes the amount of available training data as the number of non-phonetic features increases and puts a limit on the number of non-phonetic features that can be incorporated in the model. In addition, there is no data sharing between various conditions. As a consequence, the model becomes poorly estimated and the performance of the model also degrades. Secondly, the feature selection process is empirical and heuristic. Some non-phonetic features may influence only certain parts of the model and introduce very delicate changes to the model structure. For example, gender difference has a significant influence on vowels and diphtongs but not on stops or fricatives. Usually all model units are retrained based on the selected subset of data and there is no consistent theoretical framework to incorporate non-phonetic features.

In this paper, an approach of incorporating general features in high-resolution acoustic modeling is described. This approach is based on a unified maximum likelihood statistical framework. The key step in the proposed approach is a tagging scheme to tag the non-phonetic features in the training data during phonetic segmentation. The tagged non-phonetic features are used simultaneously with the regular phonetic features in the decision tree clustering process during model building. Therefore, the decision tree based model building is according to two types of knowledge sources and there is no manual separation of the training data. As a consequence, the model is built from the same set of training data regardless the number of features to which we intend to incorporate. Thus, it solves the training data depletion problem as in prior approaches and the generalized features are incorporated based on a consistent maximum likelihood framework. In this paper, we will present the application of the tagged decision tree based acoustic modeling on phonetic and non-phonetic features. We use the position of a phoneme in the word as a phonetic feature to build word position-dependent (intra- and inter-word) acoustic models. The genders of the training speakers are used as one example of non-phonetic features to improve recognition accuracy.

2. TAGGED DECISION TREE BASED ACOUSTIC MODELING

Building separate models for specific phonetic and non-phonetic features usually improves the acoustic modeling for large vocabulary speech recognition, if sufficient training data for each condition is available. While

phonetic features, like phonetic contexts, are handled by decision tree clustering [1,5,7], other features such as word positions (intra- and inter-word models) and speaker genders are not incorporated consistently. Usually they are explicitly modeled by an expansion of the phoneme or model set. The expanded set of acoustic models is usually trained by splitting the training data according to the used features and estimating separate sets of HMMs for the individual conditions. This method assumes that every state of HMMs depends uniformly on the specific conditions such as speaker's gender and the position within a word. This is obviously not correct since different phonemes may depend on a different degree to the speaker's vocal tract characteristics, such as those exhibited by the gender. Furthermore the first and last state of an HMM in inter-word position is expected to be more influenced by the word boundary than the central state.

The splitting of the training data reduces the available number of training examples for each acoustic model, which can lead to poorly estimated models. For this reason, other specific models beyond gender conditions are rarely used [8]. One solution to train specific HMMs for different features is through the use of adaptation techniques such as Maximum-A-Posteriori (MAP) adaptation [4]. In this approach, generic and condition independent models are estimated first and then adapted to the specific conditions, e.g. speaker gender. Although MAP adaptation techniques are useful, it does not change the state tying relations of the generic model. The state tying relations of the generic model may not reflect specific features in the adaptation data. Individual states in the generic model may be separated or tied together according to the likelihood estimation and occupation counts from the complete data set, which may not be optimal for the specific conditions. This situation can become acute when distributions of complete training data and adaptation data are substantially different.

For these reasons, it is preferable to use an automatic and unified approach to generate specific acoustic models for different features in a data-driven manner. In our proposed approach, this is achieved by incorporating various features at the decision tree clustering stage. Individual states for different features are only separated if this leads to a significant increase in the likelihood of the training data. The additional information about specific features is provided as tags to the decision tree clustering procedure. In our experiments gender and word position tags are used, but the concept is very general and many other non-phonetic features are applicable. The tagging procedure in our approach partitions the training data into specific subsets, based on both the phonetic context and tagged features of interest, upon which single mixture Gaussian models are trained for the untied system. Consequently, every HMM state is associated with an appropriate label, marking the specific conditions of its training data. The question set of the decision tree is extended to also include questions regarding non-phonetic features. During the construction of the decision tree the best question to split the data according to the likelihood criteria is selected. If the separation of specific conditions results in the maximum likelihood gain among all questions, separate HMM states will be constructed for the specific conditions. If no question about a particular feature is used on the path from the root node to a particular leaf, the associated tied state is used as independent of that feature. This data-driven approach prevents unnecessary data splitting and constructs a minimal set of states for a given training data and a likelihood threshold. The decision to model a feature separately is performed on the state level and thus HMMs can consist of specialized and unspecialized states for different features.

The approach of tagged decision tree is a very general one and can be used for many other features, such as speaker identity and channel conditions. In [10] several linguistic and phonetic features such as vowel stress were proposed. In our approach the tags are generalized to include any additional information, including non-phonetic features, in decision tree state tying.

3. EXPERIMENTAL RESULTS

The proposed approach of the tagged decision tree state tying was evaluated on different experiments for the Wall Street Journal (WSJ) task. 12 mel-cepstral coefficients and the normalized energy plus their 1st and 2nd order time derivatives were used as acoustic features. The cepstral mean for each sentence was calculated and removed. All HMMs have three emitting states and a left-to-right topology. Training of the acoustic parameters was based on the two-level segmental clustering algorithm described in [5]. Single mixture triphone models were estimated for all triphones exceeding a sample threshold in the training data. A minimum count threshold of 5 and 10 examples was used in our experiments, but no significant performance difference was observed. Phonetic decision tree state tying was used to cluster equivalent sets of context dependent states and to construct unseen triphones. The final triphone HMMs were built based on the tied states from the clustering. The number of mixtures for each tied state depends on the amount of training data assigned and varies from 4 to 12.

Decoding was done using a one-pass N-gram decoder [2], in which the search was conducted on a layered self-adjusting decoding graph using the cross-word triphone models. The standard SI-84 and SI-284 training data sets were used to train the WSJ models. The lexicon was

generated automatically using a general English text-to-speech system (41 phones) [3]. The language models used in the experiments were the standard trigram language models provided by NIST for the WSJ corpus.

3.1 Gender-Dependent Models

The first set of experiments is concentrated on the proposed tagged decision tree state tying for gender-dependent acoustic models. Since gender identification is not the issue of this paper, we assume genders of the test speakers to be known. Table 1 tabulates the word error rates of gender-independent (**GI**) models and different gender-dependent models trained on the WSJ SI-84 dataset.

Model		**WSJ-92**	
SI-84	#states male + female	5k-closed	20k-open
GI	3447	5.0 %	12.8 %
sGD	2633 + 2622	4.8 %	12.2 %
mGD	3447 + 3447	4.5 %	12.1 %
cGD	2753 + 2735	4.4 %	11.7 %

Table 1: Word error rates for gender-dependent acoustic models (SI-84 training data).

The first set of gender-dependent models (**sGD**) was trained by splitting the training data in a male and female subset and training completely independent HMMs for both genders. The state tying for both HMM sets is derived from two independent decision trees for the two data subsets and results in 2633 states for the male and 2622 states for the female triphones. The adaptation of gender-independent models to the male and female data using MAP results in two model sets (**mGD**) with 3447 states each, because MAP adaptation does not affect the state tying relationship in the generic model. Compared to the sGD-models the MAP adaptation provides more robust estimates for triphones and results in a small error rate reduction. The last row in Table 1 presents the results of the proposed tagged decision tree state tying approach (**cGD**). The decision tree decides, based on the data, for every state of all triphones whether the state is modeled separately for the male and female HMMs or a joint state is used. The total number of tied states (5488) is about 5% more than states in sGD-models using the same likelihood threshold. The improvement compared to the sGD-models is between 4% and 8% and compared to the gender-independent models the error rate reduction is between 9% and 12% for the two evaluation test sets.

In Table 2 the results for SI-284 trained HMMs are listed. Again the automatically clustered gender-dependent states of the cGD-model perform best. The relative error rate difference between cGD- and sGD-models is up to 9% for the 5k vocabulary. Compared to the gender-independent models the word error rate is reduced about 3% relative for both evaluation test sets.

Model		**WSJ-92**	
SI-284	#states male + female	5k-closed	20k-open
GI	8006	3.0 %	9.8 %
sGD	5835 + 6043	3.2 %	9.8 %
mGD	8006 + 8006	3.0 %	9.8 %
cGD	5984 + 6181	2.9 %	9.5 %

Table 2: Word error rates for gender-dependent acoustic models (SI-284 training data).

An analysis of the tagged decision tree indicates a phonetically reasonable behavior. States for vowels and diphtongs (except for the schwa sound, /aa/) are mostly separated for males and females, while stops and fricatives share up to 34% of their states for both genders. This appears to be consistent with the dependency of phones on vocal tract characteristics. The gender questions are competing with other questions about the phonetic context. They are used only if resulting in a maximum increase in likelihood among all questions and enough training data is available to satisfy minimum occupation requirements. A tied state (leaf in decision tree) will be shared between genders if no gender question, separating the male and female data, was used in the path from the tree root to the leaf. Phonemes with more than 20% state sharing are listed in Table 3.

t	aa	d	oy	ch	G	zh	jh	b	th	uh
34	32	30	29	25	23	22	22	22	22	21

Table 3: Sharing of states between genders for different phonemes in %.

3.2 Position-Dependent Models

The proposed tagged decision tree approach is also applied for modeling word position dependent HMMs. While some of the context dependent models (silence and noise models) occur only at word boundaries, most of the triphones appear in both inter- and intra- word positions and exhibit various degrees of dependencies on positions. Moreover, the number of occurrences of these word

position-dependent triphones in the training data also varies drastically, and some units may not have enough samples to model them separately for different word positions. Table 4 depicts the word error rates for word position- dependent HMMs (**DEP**) trained on WSJ SI-84 and SI-284 with the proposed tagged decision tree state tying approach. Results for position-independent models (**IND**) are also included for comparison.

Model		**WSJ-92**			
		5k-closed		20k-open	
		IND	**DEP**	**IND**	**DEP**
SI-84	**GI**	5.0 %	4.4 %	12.8 %	11.6 %
SI-84	**GD**	4.5 %	4.2 %	12.1 %	11.2 %
SI-284	**GI**	3.0 %	3.0 %	9.8 %	9.5 %
SI-284	**GD**	3.0 %	2.9 %	9.8 %	8.8 %

Table 4: Word error rates for word position dependent HMMs.

Using position-dependent model units leads to a 10% error rate reduction for **GI** models and a 5% reduction for **GD** models compared to the baseline results. For SI-84 system, the total number of tied states with position-dependent HMMs increased about 30% from approx. 3400 to 4400. For the SI-284 system, a small error rate reduction can be achieved for the gender-independent 20k task and a significant 10% word error reduction for the gender-dependent models. Table 5 illustrates the rate of state sharing between inter- and intra- word models for different phonetic classes.

Fricatives	Stops	Nasals	Vowels
65 %	71 %	76 %	80 %

Table 5: Sharing of states between word position dependent models for different phoneme classes.

The sharing of states between word position-dependent models varies from 30% for the /dh/ sound and 100% for rare phones with a small number of decision tree leaves like /zh/. These rare phonemes don't have sufficient examples in the training data to allow a state split into position-dependent variants of the same phoneme. Some vowels like /eh/ are not much affected by their position in a word and share up to 88% of the states. The tagged decision tree clustering automatically balances the need to generate separate position-dependent states against the availability of training data and prevents unnecessary data splits.

4. SUMMARY

In this paper, an approach of acoustic modeling based on tagged decision tree clustering is described. In this approach, a tagging process is used to include non-phonetic features into the state clustering. It has two unique advantages. First, both phonetic and non-phonetic features can be incorporated in the decision tree tying process based on a consistent maximum likelihood framework. Second, it is totally data driven and is based on the full training data. Therefore, it alleviates the problem of training data depletion encountered in condition dependent acoustic modeling. Experimental results on large vocabulary speech recognition tasks indicate that up to 10% word error rate reduction can be achieved with the proposed approach.

5. REFERENCES

[1] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree Based State Tying for High Accuracy Modeling", ARPA Workshop on Human Language Technology, Princeton, NJ, Morgan Kaufmann Publishers, March 1994.

[2] Q. Zhou, and W. Chou, "An Approach to Continuous Speech Recognition Based on Layered Self-Adjusting Decoding Graph", ICASSP 97, Munich, Germany, April 1997.

[3] R.W. Sproat and J.P. Olive, "Text-to-Speech Synthesis", AT&T Tech. Journal, 74, pp. 35-44, 1995.

[4] J.-L. Gauvain, C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, 1994.

[5] W. Reichl and W. Chou, "Decision Tree State Tying Based on Segmental Clustering for Acoustic Modeling" ICASSP 98, Seattle, May 1998.

[6] D.B. Paul, "Extensions to Phone-State Decision-Tree Clustering: Single Tree and Tagged Clustering", ICASSP 97, Munich, Germany, April 1997.

[7] W. Chou and W. Reichl, "High Resolution Decision Tree Based Acoustic Modeling Beyond CART", ICSLP 98, Sydney, Australia, November 1998.

[8] C. Neti, S. Roukos, "Phone-Context Specific Gender-Dependent Acoustic-Models for Continuous Speech Recognition, IEEE Automatic Speech Recognition and Understanding Workshop, Santa Barbara, December 1997.