Accessing Language Specific Linguistic Information for Triphone Model Generation: Feature Tables in a Speech Recognition System

Supphanat Kanokphara, Anja Geumann, and Julie Carson-Berndsen

Department of Computer Science
University College Dublin
Ireland
{supphanat.kanokphara, anja.geumann, julie.berndsen}@ucd.ie

Abstract

This paper is concerned with a novel methodology for generating phonetic questions used in tree-based state tying for speech recognition. In order to implement a speech recognition system, language-dependent knowledge which goes beyond annotated material is usually required. The approach presented here generates phonetic questions for decision trees are based on a feature table that summarizes the articulatory characteristics of each sound. On the one hand, this method allows better language-specific triphone models to be defined given only a feature-table as linguistic input. On the other hand, the feature-table approach facilitates efficient definition of triphone models for other languages since again only a feature table for this language is required. The approach is exemplified with speech recognition systems for English and Thai.

Introduction

One of the main advantages of statistical speech recognition systems is that they are assumed to not to require a lot of language-specific linguistic knowledge; once an annotated corpus is available as acoustic training data, the system can be trained to build models from that data. However, since most current speech recognition systems are based on context-dependent Hidden Markov models (HMM), this requires a large number of context-dependent units to be trained. Unfortunately no single corpus (or even multiple corpora) can possibly contain such a large number of units.

In order to alleviate this problem and strike a balance between the number of context-dependent units and the limited acoustic training data, tree-based state tying is commonly employed (Odell, 1995) which allows parameters which exhibit similarity to be shared between context-dependent units. Traditionally this approach involves only with left and right context-dependency is concerned, i.e. with context-dependent units known as triphones. The level of similarity is determined automatically from phonetic decision trees.

Phonetic decision trees are composed of a set of phonetic questions. These questions aim to determine the similarity of the contexts and often rely on the phonetic judgments of a human expert who can determine whether the contexts refer to similar contexts based on phonetic categories such as *consonant*, *vowel* or *labial*. In order to reduce the manual effort in the question construction procedure, there have been a number of attempts to automatically generate questions for tree-based state tying systems (Beulen & Ney, 1998), (Singh, Raj & Stern, 1999), (Chelba & Morton, 2002) and (Diehl & Moreno, 2004). These data-driven automatic question generation systems ignore explicit phonetic knowledge and rely purely on acoustic training data. Similar acoustic models are then used as a substitute context within triphones. Such data-driven systems are, in the majority of cases, demonstrably as good as, or even better than, manually generated questions. However, the systems appear to under-perform in three specific cases: when the data is insufficient, when the phonetic units are poorly described, and when the training conditions are noisy. The work presented in this paper aims to overcome such shortcomings.

Generating questions from a symbolic description which summarizes the phonetic characteristics of a particular language is both robust to poor recording quality of the acoustic material and does not rely on a very large corpus. The greatest disadvantage of this approach so far has been that it is quite labour intensive and time consuming. A more elegant and efficient approach is to generate questions for the decision trees from phonetic features as defined in a feature table.

Feature-table-based systems have two main advantages. The first advantage is that feature tables are commonly used in phonological descriptions of languages and are thus readily available. For this reason, the system can be extended to a new language easily even without a lot of language specific linguistic knowledge on the part of the developer. Even if no feature table exists for some languages (i.e. in the case of lesser studies languages), the feature-table-based system limits the manual aspects of the process to feature table construction by a human expert where previously it was necessary for the expert to evaluate the contexts of all units to be modeled. Furthermore, the feature table, once constructed by a human expert, can be re-used not only for other speech recognition systems, but also for more experimental phonetic and phonological studies. In the manual construction of phonetic questions, units are grouped according to features and combinations of features (co-occurrences). Thus every possible feature combination has to be considered. In a feature-table-based system, on the other hand, human effort is reduced to tagging each unit with features if there no feature table already exists. The second advantage of feature-table-based

	voice	manner	place	v_height	round	tense
/b/	voiced	stop	labial	nil	nil	nil
/p/	voiceless	stop	labial	nil	nil	nil
/m/	voiced	nasal	labial	nil	nil	nil
/u/	voiced	vowel	back	high	+round	tense
/i/	voiced	vowel	front	high	-round	tense

Table 1: Sample entries from the English feature table.

systems is that questions can easily be modified and adjusted in order to achieve higher system accuracy, since the topology of a feature table is less complex than that of a phonetic question set. Modifications can be made to a feature table and the set of questions can be generated anew. Feature-table-based question generation makes more transparent, which parameters are useful in question generation, which feature co-occurrences are suitable and which features and combinations do not contribute to higher recognition accuracy (Kanokphara & Carson-Berndsen, 2004). Furthermore, the feature-table-based approach simplifies system development not only for a new language but also when different phone sets are designed and tested in order to find the best phone set for a speech recognition system (Kanokphara & Carson-Berndsen , submitted).

Tied-State Triphone Models and Features

In general, a context-dependent speech recognition system consists of two phases: building context-independent HMMs and from these constructing tied-state triphones. Two inputs are required for the training procedure, namely an annotated speech corpus (i.e. speech signal and corresponding transcription files) and a set of phonetic questions which allows the independent HMMs to be clustered into triphones. While the annotated speech corpus is required for the procedure as a whole, the phonetic questions are essential only for creating tied-state triphone models.

The quality of the annotation of a speech corpus is crucial for a speech recognition system. Similarly, the quality of the phonetic questions is important for creating tied-state triphone models. In this paper, do not address issues of corpus quality, i.e. we assume that the quality of the corpus is ideal, rather we concentrate on how to improve the phonetic question construction process. Our goal here is in line with the aims of the triphone clustering described in Netsch & Bernard (2004) namely to economize on time and human resources. However, the approach and practical conditions are different. Netsch & Bernard (2004) intend to use two feature tables (from a reference and a target language) to implement a speech recognition system for a target language where both speech corpus annotation and phonetic question quality are poor. Our system focuses on simplifying the question generation process. Therefore, the quality of the acoustic data and the speech annotation can be largely ignored which is not the case with a data-driven automatic question generation methods.

In triphone clustering, (Netsch & Bernard, 2004) defined two feature tables and mapped universal questions to target language questions. This limits a study of feature co-occurrences to universal questions but does not allow for language-specific questions. In contrast, the approach described in the next section, generates questions from a language-specific feature table and requires no universal questions. As a result, there is no necessity to construct universal questions manually and adaptation can be performed using language-specific questions.

Feature Tables

In general, a feature table consists of a phonetic segment (phone) with and associated set of phonetic features. Each feature belongs to a *tier* which is represented as a column in the table. That is to say some features are mutually exclusive, and therefore not all combinations of features exist and therefore features are grouped on a tier according to features which cannot co-occur. The feature tables underlying this approach are not restricted to articulatory features. Feature tables can contain acoustic information and also include other features like syllable position, gender, etc. In Tables 1 and 2 we present sample entries of the English and Thai feature sets used in the study presented here. Due to space limitations, the complete feature tables are not included in the paper but are available from the authors on request. Note that the features used in each table differ. Furthermore, the Thai feature table contains a greater number of features and clustering segments together to some degree. Additional features are included which refer to the position of the segment within the syllable since a number of consonants have a different pronunciation in syllable final position (e.g. /□/ indicates that the preceding consonant is unreleased). The differences among the two feature systems describe the wide range of potential feature table inputs. As it stands, the feature set used for English is phoneme based and an example of a basic feature table that can be easily constructed for any language without in depth knowledge about its phonotactics (sound patterns) and yet still produce a workable number of phonetic questions for a speech recognition system. Since English has a more complex phonotactic structure, it would not have been practical to integrate all syllable position information into this table. Phonotactic constraints which are useful for syllabification

	Position	Consonant	Length	Voice	Stop	Manner	Place	Static	Round	Height
b	onset	consonant	single	voiced	stop	voiced-stop	labial	nil	nil	nil
bl	onset	consonant	cluster	voiced	stop:non	voiced-stop:lateral	labial:alveolar	nil	nil	nil
br	onset	consonant	cluster	voiced	stop:non	voiced-stop:trill	labial:alveolar	nil	nil	nil
i	nucleus	non-con	short	voiced	nil	vocalic	front	static	unround	high
ia	nucleus	non-con	short	voiced	nil	vocalic	front:central	non	unround	high:low
ii	nucleus	non-con	long	voiced	nil	vocalic	front	static	unround	high
iia	nucleus	non-con	long	voiced	nil	vocalic	front:central	non	unround	high:low
m	onset	consonant	single	voiced	non	nasal	labial	nil	nil	nil
m□	coda	consonant	single	voiced	non	nasal	labial	nil	nil	nil
p	onset	consonant	single	unvoiced	stop	unaspirated	labial	nil	nil	nil
p□	coda	consonant	single	unvoiced	stop	unaspirated	labial	nil	nil	nil
ph	onset	consonant	single	unvoiced	stop	aspirated	labial	nil	nil	nil
phl	onset	consonant	cluster	unvoiced:voiced	stop:non	aspirated:lateral	labial:alveolar	nil	nil	nil
phr	onset	consonant	cluster	unvoiced:voiced	stop:non	aspirated:trill	labial:alveolar	nil	nil	nil
pl	onset	consonant	cluster	unvoiced:voiced	stop:non	unaspirated:lateral	labial:alveolar	nil	nil	nil
pr	onset	consonant	cluster	unvoiced:voiced	stop:non	unaspirated:trill	labial:alveolar	nil	nil	nil
u	nucleus	non-con	short	voiced	nil	vocalic	back	static	round	high
uu	nucleus	non-con	long	voiced	nil	vocalic	back	static	round	high
uua	nucleus	non-con	long	voiced	nil	vocalic	back:central	non	round:unround	high:low

Table 2: Sample entries from the Thai feature table.

at a later stage, can be modeled separately (Carson-Berndsen, 1998).

For each of the feature tables, the actual question generation algorithm remains unchanged, only the feature table differs. The approach is thus language-independent. We now present some further details of how phonetic questions are generated from such tables.

Feature-Table-Based Automatic Question Generation System

The algorithm used for automatic question generation is the same as in (Kanokphara & Carson-Berndsen, 2004). However, we summarize the algorithm briefly here and then concentrate on examples of questions which are generated using this method. Some results, which highlight the contribution of this approach to the performance of speech recognition engines for English and Thai, are presented.

The algorithm contains the following steps:

1. Read the segments and their features, i.e. each row of the feature table.
2. List all features with the sounds they specify, e.g. from Table 1 *voiced* {/b/, /m/, /u/, /i/}, *voiceless* {/p/}, *stop* {/b/, p/} etc.
3. Combine features from different tiers with "and", corresponding to set intersection, e.g. "*voiced* and *stop*" {/b/, /m/, /u/, /i/} ∩ {/b/, p/} to give {/b/}; "*voiceless* and *vowel*" {/p/} ∩ {/u/, /i/} to give {}.
4. Delete redundant features, e.g. the feature *back* and the feature +*round* denote the same sound {/u/} in Table 1, so +*round* can be omitted.
5. Delete empty sets and those combinations that contain the *nil* feature.
6. Generate questions.

The strategy for creating feature co-occurrence questions relies on the fact that there is no identical feature on different tiers except "nil". for every cross-tier feature co-occurrence, if there are one or more units corresponding to the co-occurrence, this co-occurrence is used as a question. With this assumption, impossible questions are automatically discarded. For example, according to the English feature table, there will never be a question which asks whether a particular context is *voiceless* and *vocalic* question because these features do not co-occur on; the intersection of the segments which have the features voiceless and vocalic will thus yield the empty set.

Typical phonetic questions for the decision tree generated using the English feature table are:

- "Does the left the context of triphone X possess a *stop* feature?"
- "Does the right the context of triphone Y possess a *stop* and a *voiced* feature?"

Based on the sample entries given in Table 1 only, the first question is equivalent to a set of questions asking whether the left context of triphone X is one of {/p/, /b/}; the second question is equivalent to a set of questions asking whether the right context of triphone Y is /b/ (i.e. the intersection of *stop* {/b/, /p/} and *voiced* {/b/, /m/, /u/, /i/}). The full feature table for English results in a set of 238 such questions (see Kanokphara & Carson-Berndsen, submitted).

Typical phonetic questions generated using the Thai feature table are:

- "Does the left the context of triphone X possess a *unvoiced:voiced* cluster?"
- Does the left portion of the right context of triphone Y possess a *front* feature?

Based on the sample entries presented in Table 2 only, the first question is equivalent to a set of questions asking whether the left context of triphone X is one of {/ph/, /phr/, /pl/, /plr/}; the second question is equivalent to a set of questions asking whether the right context of triphone Y begins with {/ia/, /iia/}, /i/, /ii/}. The full feature table for Thai results in a set of 1046 such questions (see Kanokphara & Carson-Berndsen, 2004). The questions thus generated are then used by the tree-based state tying technique to determine which triphone models can be grouped together or clustered because they share the same or a similar context.

The full set of phonetic questions for English and the set of phonetic questions for Thai were then used by as the basis for generating tied-state models during the second phase of speech recognition using HTK. Table 3 presents the results achieved by the system developed on this basis. The English results were obtained using the TIMIT corpus and the Thai results using the NECTEC-ATR Thai corpus. The results compare favourably with other speech recognition systems for these languages.

	English	**Thai**
% WER	28.86	21
No. of Questions	238	1046

Table 3: WER for English and Thai Systems

Conclusion

This paper has presented a novel methodology for generating phonetic questions for use in tree-based state tying for speech recognition. In this way, phonetic information can be incorporated explicitly into HMM-based context-dependent speech recognition systems. Tree-based state tying requires a decision to be made as to which left and right contexts of triphone models together based on phonetic similarity which is automatically extracted from a phonetic feature table. Traditionally such question sets were constructed manually based on similarity judgments of a human expert. The feature-table-based approach described in this paper generates phonetic questions from a feature table automatically instead thus overcoming the shortcomings of other data driven techniques which rely on large quantities of good quality acoustic data. For our approach only a feature-table is require for each language, which if not readily available, is easy to produce for any given language since feature descriptions are commonly used in phonetic and phonological studies. Furthermore, such tables can be expanded to include additional features as highlighted above for the Thai feature table or can be kept basic as the English feature table.

Future work is concerned with evaluation of other feature tables, investigating the contributions of different types of features (acoustic, articulatory, position, gender etc.) to the recognition process (see Geumann 2004). The feature-table-based system is proving to be very convenient where a speech recognition system has to be developed for a new language or where a phone set modification is required in order to improve performance.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland for the support under Grant No. 02/IN1/I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. The authors would like to thank NECTEC for the use of NECTEC-ATR Thai corpus.

References

Beulen K. and Ney H. (1998): Automatic Question Generation for Decision Tree Based State Tying. Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 805-809.

Carson-Berndsen, J. (1998): *Time Map Phonology*, Kluwer, Dordrecht.

Chelba C. and Morton R. (2002): Mutual Information Phone Clustering for Decision Tree Induction. *Proc. Int. Conf. on Spoken Language Processing (ICSLP).*

Diehl F. and Moreno A. (2004): Acoustic Phonetic Modelling using Local Codebook Features. *Proc. Int. Conf. on Spoken Language Processing (ICSLP).*

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L. (1993): *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*, NIST.

Geumann, A. (2004): Towards a New Level of Annotation Detail of Multilingual Speech Corpora. *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 1096-1099.

Kanokphara, S. and Carson-Berndsen J. (2004): Automatic Question Generation for HMM State Tying using a Feature Table. *Proc. Australian Int. Conf. on Speech Science & Technology (ASST).*

Kanokphara, S. and Carson-Berndsen J. (submitted): *Feature-Table-Based Automatic Question Generation for Tree-Based State Tying: A Practical Implementation,* submitted to the 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Bari, Italy.

Netsch, L., Bernard, A. (2004): Automatic and language independent triphone training using phonetic tables, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada.

Odell, J.J. (1995): *The Use of Context in Large Vocabulary Speech Recognition.* Ph.D. diss., Cambridge University, Cambridge.

Singh, R., Raj, B. and Stern, R. M. (1999): Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models. *Proc. International Conference on Spoken Language Processing (ICSLP)*, vol. 1, 117-120.