Better HMM-Based Articulatory Feature Extraction with Context-Dependent Model

Supphanat Kanokphara and Julie Carson-Berndsen

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

Abstract

The majority of speech recognition systems today commonly use Hidden Markov Models (HMMs) as acoustic models in systems since they can powerfully train and map a speech utterance into a sequence of units. Such systems perform even better if the units are context-dependent. Analogously, when HMM techniques are applied to the problem of articulatory feature extraction, context-dependent articulatory features should definitely yield a better result. This paper shows a possible strategy to extend a typical HMM-based articulatory feature extraction system into a context-dependent version which exhibits higher accuracy.

Introduction

HMMs are predominantly used as acoustic models in most current speech recognition systems. The reason for this is that HMMs can normalize the time-variation of the speech signal and characterize speech signal statistically in the optimal sense. Due to a large number of vocabularies in the real world, it has been most practical to design acoustic models at the phonetic level. However, this has the drawback effect that phonetic units representing parts of speech are not easily designed as a result of co-articulation. This problem can be alleviated by ensuring that the units are context-dependent.

The context-dependent HMM-based approach has proven to be fruitful in recent speech recognition systems. However, by completely ignoring linguistic knowledge and relying only on statistical models like HMMs, the systems can achieve only a certain level of success (Lee, 2004). One of the problems of stochastic systems is that they are too restrictive and thus not fully applicable in adverse environments (noisy, out-of-task, out-of-vocabulary, etc). Many researchers aware of this problem and are trying to integrate more explicitly knowledge-based and statistical approaches.

Integrating articulatory features into the system is one of the ways in which such a hybrid system can be achieved (Carson-Berndsen, 1998), (Kirchhoff, Fink and Sagerer, 2002) and (Richardson, Bilmes and Diorio, 2003). With this approach, systems are better modeled by means of linguistic knowledge and hence yield better recognition results.

In order to use articulatory features in for speech recognition, many articulatory feature extraction systems has been developed (Abu-Amer and Carson-Berndsen, 2003), (Chang, Greenberg and Wester, 2001) and (Ali et al. 1999). Among these, the HMM-based system seems to outperform the others. We therefore set this HMM-based system as our baseline.

HMM-based speech recognition systems are usually upgraded by using context-dependent phones (Odell, 1995). Similarly, we choose to upgrade HMM-based articulatory feature extraction systems with context-dependent features. In this paper, we start by investigating possible ways to broaden context-independent units into context-dependent counterparts. We then discuss a normal HMM-based articulatory feature extraction system and suggest an appropriate way to model context-dependent features. The context-dependent and context-independent systems are then compared via experiments. Finally, conclusions are drawn and future directions discussed.

Context-Dependent Units for Speech Recognition

One of the difficulties that is most discussed when context-dependent units are introduced to speech recognition systems is striking a balance between the level of information in models and the limited amount of acoustic training data. This is because the number of context-dependent units is naturally large. To illustrate this, let “a” be a context-independent unit. Then, its context-dependent version can be labeled as “b-a+c” where “b” and “c” are preceding and succeeding units, respectively. Therefore, if

N is the number of context-independent units, the number of context-dependent counterparts will be NxNxN which is unacceptable for training.

In this paper, three strategies for making context-dependent units trainable are presented, namely, backing-off, smoothing and sharing. These techniques require some algorithms to determine what parameters underlie backing-off, smoothing and sharing, respectively.

Backing-Off

This is the first and simplest strategy for training context-dependent units. When there is insufficient data for training a model, that model should back-off and some less informative but trainable model should be used instead. For example, if a triphone has only a few examples in the training data, a biphone should be used. If a biphone is still not trainable, monophone should be used. With this strategy, it is possible to insure that all models are well trained. The disadvantage of this strategy, however, is that the difference between more and less informative models is too large when a backing-off occurs.

Smoothing

In 1989, Lee and Hon proposed an alternative way to keep a balance between information in models and sufficiency of training data. This method uses interpolation between less informative but trainable and more informative but un-trainable models. The advantage of this strategy is that it can smooth deeply into the state level, in contrast to the backing-off strategy which is applied only at the model level.

Sharing

The sharing strategy is perhaps the most frequently used for balancing trainability and information of models. Sharing schemes can be divided into two approaches, namely bottom-up and top-down approaches. The bottom-up approach starts by generating all context-dependent units occurring in the training data. Some algorithm then is used to find similar states and tie them together. In this way, tied states use the same training data and make the system trainable. However, some examples are required for searching similar states. This makes defining good unseen models impossible. The top-down approach, on the other hand, uses linguistic knowledge to form a decision tree. This tree then is used to cluster and tie states hierarchically. This tree can also synthesize unseen models linguistically and therefore it does not suffer from the same problem as the bottom-up approach (Odell, 1995).

HMM-Based Articulatory Feature Extraction System

The HMM, by design, is used to map some uncertainty signal into a sequence of units. These units can be words, syllables, demi-syllables, phones, etc. The articulatory feature extraction presented here also uses this type of HMMs to map a speech signal into a sequence of features. In the linguistic sense, this integrates articulatory information into a statistical system and thus results in a better system. Moreover, as each HMM-based system recognizes a sequence of features on each tier independently, it allows overlap among features on each tier. This means that sequences of features from the system also capture coarticulation phenomena. In the statistical sense, the number of classes for a system to recognize is reduced and hence more robust models can be built.

System Overview

Before performing articulatory feature extraction, a feature table listing different features on each tier has to be properly assigned. In this paper, we follow the same feature table as in (Abu-Amer and Carson-Berndsen, 2003). The feature table contains 6 different tiers: *manner*, *place*, *voicing*, *vowel type*, *vowel height* and *lip rounding*.

The articulatory feature extraction system in this paper is constructed using HTK (http://htk.eng.cam.ac.uk/) which is now widely used for HMM based speech recognition experiments. The acoustic model training system starts by converting the speech signal into a sequence of vector parameters with a fixed 25 ms frame and a frame rate of 10 ms. Each parameter is then pre-emphasized with the filter P(z) = 1-0.9*z^(-1). The discontinuities at the frame edges are attenuated by using Hamming window. A fast Fourier transform is used to convert time domain frames into frequency domain spectra. These spectra are averaged into 24 triangular bins arranged at equal mel-frequency intervals (where $f_{mel} = 2595\log_{10}(1+f/700)$). f denotes frequency in Hz. 12 dimensional mel-frequency cepstral coefficients (MFCCs) are then obtained from cosine transformation and lifter. The normalized log energy is also added as the 13^{th} front-end parameter. The actual acoustic energy in each frame is calculated and the maximum found. All log energies are then normalized with respect to maximum and log energies below a silence floor (set to -50 dB) clamped to that floor.

These 13 front-end parameters are expanded to 39 front-end parameters by appending first and second order differences of the static coefficients. The chosen parameters chosen have been used extensively (Davis and Mermelstein, 1980) and have proven to be one of the best choices for HMM-based speech recognition systems.

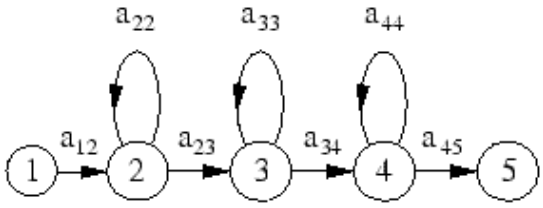

Figure 1 5-state left-right HMM

Flat start training is then used for model initialization according to features on each tier. Flat start training is a training strategy from HTK requiring no time-annotated training transcriptions for model initialization. Each model contains 5 states and the covariance matrices of all states are diagonal. Figure 1 shows a 5-state left-right HMM as used in the system.

Maximum likelihood estimators are used to train HMM parameters (Juang, 1985). The number of training iterations after each change is determined automatically in line with (Tarsaku and Kanokphara, 2002). The models are finally expanded to 15 mixtures except for the *manner* tier where the models are expanded to only 5 mixtures. These mixture numbers are the same as those used by (Abu-Amer and Carson Berndsen, 2003).

The language model is trained from the training set on each tier using back-off bigram. The language model provides feature constraints which correspond to the intra-feature-model transition probabilities. For the recognition process, the Viterbi algorithm is used without any pruning factor.

TIMIT Corpus

The standard TIMIT corpus (Garofolo et al. 1993) consists of 3600 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the U.S., of which 462 are in training set and 168 are in the testing set. There is no overlap between the training are testing sentences, except 2 dialect (SA) sentences which were read by all speakers. The training set contains 4620 utterances and the testing set 1680 (112 males and 56 females). The core test set, which is the abridged version of the complete testing set, consists of 192 utterances, 8 from each of 24 speakers (2 males and 1 female from each dialect region). Normally, SA sentences are eliminated from the training/testing set because they occur in both training and testing set.

All utterances were recorded in a noise-isolated recording booth. The speech was directly digitized at a sample rate of 20 kHz with the anti-aliasing filter at 10 kHz. The speech was then digitally filtered, debiased and downsampled to 16 kHz.

The training/testing set in this paper is exactly the same set as in (Abu-Amer and Carson-Berndsen, 2003) which is full set training (4620 utterances) and testing set without SA (1344 utterances). All TIMIT phonemic training and testing transcriptions are transformed to feature transcriptions automatically.

Context-Dependent Articulatory Feature Extraction System

According to the discussion of the many possible techniques for making context-dependent units trainable above, the top-down approach would seem to be the best way of balancing the parameters of the model. Unfortunately, this technique requires some linguistic knowledge in terms of phonetic questions which are in fact a list of unit classes defined with respect to unit features. For context-dependent units in speech recognition, units are usually phones or larger units in which there are always some mutual feature characteristics which can be shared. However, in our system, we want to construct context-dependent features which have no further mutual feature characteristics for sharing like phones or larger units.

Since the top-down approach is not appropriate for our context-dependent features, we start to investigate other strategies. The bottom-up approach is discarded as it is weak for unseen context-dependent units. In this paper, backing-off approach is chosen even though the difference between more and less informative models is too large and if data is very sparse, there will be too many context-independent units in the system.

There are at least two reasons why backing-off technique is still selected for our context-dependent features. Firstly, the number of feature classes required by the system is definitely less than the number of phones or larger unit classes. Therefore, the problem of too many context-independent units in the system should not arise. Secondly, this technique is more easily extensible than the smoothing technique to obtain more accurate models. For example, Lamel and Gauvain (1993) proposed gender-dependent models as an extension to backed-off context-dependent models for gender-independent speech recognition. This paper showed a better result than (Lee and Hon, 1989) using the smoothing technique.

A context-dependent articulatory feature extraction system can be easily constructed from a single mixture context-independent system. The context network expansion is cross-word. The training algorithm is the same as for context-independent system except that the training transcription is changed to be context-dependent version. After the models are trained, the number of mixtures is then again increased to 15 except for *manner* tier which is increased to only 5.

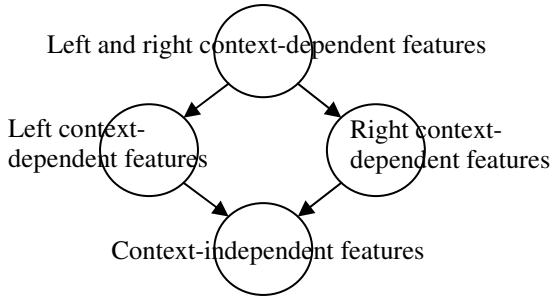

Figure 2 Backing-off hierarchy

Left, right, or left and right context dependency is determined automatically according to their frequencies in the training transcription. In this paper, when left and right context dependent feature frequency is less than 40, the highest frequency feature between left and right context dependent features is used. If both left and right context dependent feature frequency is less than 40, the context-independent feature is used. Figure 2 illustrates backing-off hierarchy.

System Performance

Percentage correct and accuracy of recognized and reference feature sequences are the measures used to evaluate system performance. These can be found by matching each of the recognized and reference label sequences by performing an optimal string match using dynamic programming. Once the optimal alignment has been found, the number of features, substitution features, deletion features and insertion features are counted and calculated. The difference between percentage correct and accuracy is that percentage correct ignores insertion features while percentage accuracy does not.

Context-Independent Articulatory Feature Extraction System

In this subsection, we compare our context-independent system with (Abu-Amer and Carson-Berndsen, 2003). Table 1 shows the (Abu-Amer and Carson-Berndsen, 2003) result and table 2 shows our result. The comparison of the two tables indicates that our system performs better than (Abu-Amer and Carson-Berndsen, 2003). According to the tables, only %correct on manner tier and %accuracy on vowel type tier exhibit worse results. The better results in table 2 are highlighted by using bold italic font type.

Context-Dependent Articulatory Feature Extraction System

The result from our experiment is very promising. The context-dependent system gives the result as shown in table 3. Both %correct and %accuracy of all tiers yield better results than the results in table 2. Our context-dependent system is also better than the system from (Abu-Amer and Carson-Berndsen, 2003) on every tier.

	place	manner	vowel height	vowel type	round	voice
% correct	75.79	88.68	83.28	87.02	88.71	70.9
% accuracy	59.51	66.6	66.94	69.5	70.08	61.74

Table 1 Result from (Abu-Amer and Carson-Berndsen, 2003)

	place	manner	vowel height	vowel type	round	voice
% correct	***77.13***	87.67	***84.64***	***91.41***	***90.75***	***96.97***
% accuracy	***68.15***	***75.94***	***76.4***	64.82	***80.23***	***64.46***

Table 2 Result from our context-independent system

	place	manner	vowel height	vowel type	round	voice
% correct	78.48	88.8	86.08	93.62	92.51	97.12
% accuracy	73.36	77.37	80.03	68.08	86.08	72.94
No. of models	561	121	91	48	45	18

Table 3 Result from our context-dependent system

The number of models for the *place* tier expands from 9 to 561, for the *manner* tier from 6 to 121, for the *vowel height* tier from 5 to 91, for the *vowel type* tier from 4 to 48, for the *lip rounding* tier from 4 to 45., and for the *voicing tier* from 3 to 18. According to these numbers, we can see that most of the models are context-dependent. For example, for the *place* tier, the total number of possible left and right context-dependent features are 9x9x9 which is 729. This means that only 23% of models for the *place* tier are not left and right context-dependent features.

Conclusion

There are many aspects of this research worth emphasizing here. Firstly, to implement better speech recognition systems, hybrid approaches which use statistical and linguistic knowledge are very attractive. Applying articulatory features is one of the possible ways to integrate linguistic knowledge into stochastic speech recognition systems. In order to build good articulatory-feature-based

systems, reliable articulatory feature extraction systems have to be investigated. In this paper, we proposed an alternative way to efficiently extract articulatory features from speech utterances.

Secondly, as articulatory features are common to most languages, this makes our system language-independent, although clearly the set of features does have to be extended beyond the set used in this paper (see Geumann, 2004).

Thirdly, as our system is based on HMM, many useful techniques for HMM-based speech recognition systems can also be applied to our system. However, articulatory feature extraction and phone (or larger unit) recognition systems cannot be treated in exactly the same way. Some technique has to be customized for articulatory feature training. For example, in this paper, in order to use a context-dependent technique, the back-off approach is preferred above the top-down approach.

Fourthly, our system can also be extended further by integrating more linguistic knowledge into our system. For example, on the *voice* tier, if some segment of the speech utterance is recognized as *unvoiced*, it cannot be recognized to be *vocalic* on the *manner* tier.

Finally, from our experiment, the context-dependent articulatory feature extraction system proves very promising for *vowel height* and *lip rounding* tiers in particular. On those tiers, more than 80% for both correction and accuracy can be achieved.

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.

References

Lee, C. 2004. From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition. In Conference Program and Abstract Book, Eighth International Conference on Spoken Language Processing, 109-112. Jeju Island, Korea.

Carson-Berndsen, J., 1998. Time Map Phonology: Finite State Models and Event Logics in Speech Recognition. Kluwer Academic Publisher, Dordrecht.

Kirchhoff, K., Fink, A., G. and Sagerer, G. 2002. Combining Acoustic and Articulatory Feature Information for Robust Speech Recognition. Speech Communication 37: 303-319.

Richardson, M., Bilmes, J., and Diorio, C. 2003. Hidden-Articulator Markov Models for Speech Recognition. Speech Communication 41: 511-529.

Abu-Amer T. and Carson-Berndsen J. 2003. HARTFEX: A Multi-Dimensional System of HMM Based Recognizers for Articulatory Feature Extraction, In Proc. Eurospeech. Geneva, Switzerland.

Chang, S; Greenberg, S. and Wester, M. 2001. An Elitist Approach to Articulatory-Acoustic Feature Classification, In Proc. Eurospeech, Aalborg.

Ali, A. M. A., Van der Spiegel, J., Mueller, P., Haentjaents, G. and Berman, J. 1999. An Acoustic-Phonetic Feature-Based System for Automatic Phoneme Recognition in Continuous Speech, In Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS), 118-121.

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

Lee, K.F., Hon, H.W. 1989. Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Trans. Acoustic Speech and Signal Processing* 37(11).

HTK Speech Recognition Toolkit, http://htk.eng.cam.ac.uk/, Cambridge University, Engineering Department.

Davis, S.B., Mermelstein, P. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoustic Speech and Signal Processing* 28(4). 357-366.

Juang, B.H. 1985. Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains. *AT&T Tech. J.* 64(6).

Tarsaku, P., Kanokphara, S. 2002. A Study of HMM-Based Automatic Segmentations for Thai Continuous Speech Recognition System. In Proc. Symposium on Natural Language Processing, 217-220. Thailand.

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 CD-ROM, NIST.

Lamel, L.F., Gauvain, J.L. 1993. High Performance Speaker-Independent Phone Recognition Using CDHMM. In proc. In Proc. EuroSpeech, 121-124. Berlin.

Geumann, A. 2004. Towards a New Level of Annotation Detail of Multilingual Speech Corpora. In Proc. Int. Conf. Spoken Language Processing, Jeju Island, Korea.