Toward an Optimum Feature Set and HMM Model Parameters for Automatic Phonetic Alignment of Spontaneous Speech

Montri Karnjanadecha[1] and Stephen A. Zahorian[2]

[1] Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University
Hat Yai, Songkhla 90112, Thailand
[2] Department of Electrical and Computer Engineering, Binghamton University
Binghamton, NY 13902, USA
montri@coe.psu.ac.th, zahorian@binghamton.edu

Abstract

Many speech segmentation techniques have been proposed to automate phonetic alignment. Most of the techniques require, however, labeled data to train, and perform well only for read, high-quality speech. Automatic phonetic alignment, for lower quality varied data with no labeled training data, the subject of this paper, is a much more challenging domain. An HMM-based automatic speech recognizer was used in this study to determine phonetic sequences and boundaries of "open source" speech data, retrieved from public websites. The HMM models were initially trained using the TIMIT database and subsequently adapted to each passage. Standard frontend features such as MFCC, LPCC and PLP, and features computed by applying the DCT directly to the short-time spectrum (DCTC) were evaluated using TIMIT data. The "best" parameter set was found to be DCTC_78 and these parameters were used to align the speech data of interest.

Index Terms— speech segmentation, phonetic alignment, speech recognition

1. Introduction

Although automatic speech recognition (ASR), automatic speaker identification, and automatic language identification have all improved dramatically, there are still large performance gaps between performance obtained with clean speech and speech spoken and recorded in more natural environments. The Open Source Multi-Language Audio (OSMLA) database project was initiated to collect real-world speech data available in audio/video format from public websites and provide labeling at various resolutions including phonetic level transcriptions [1]. The difficulty with speech database development is the requirement for human input in the annotation and labeling process. Ideally, each recording should be listened to and evaluated by a human listener. This is highly problematic for collecting a large speech database, especially when time-aligned phonetic transcriptions are desired.

Time aligned transcriptions of speech data are, however, very useful for speech research. For example, a speech recognizer based on Hidden Markov Models (HMMs) ideally should have access to phonetic boundaries of every phone in the training data for model initialization. TIMIT [2] is one of the most widely-used speech corpora for the speech research community over the past 2 decades. The labeling information provided with this corpus has greatly contributed to its popularity and continual use, despite its small size (~5.38 hours of read speech).

Several automatic methods for speech segmentation have been investigated [3,4,5,6,7]. Some of the automatic techniques, such as those reported in [3, 4], have yielded segmentation accuracy comparable to that from manual methods. However, all of these techniques were for read good quality speech, with ground truth information available for training and testing. Many published works focus on speech segmentation techniques for speech synthesis applications [5,8,9], again using high quality speech. In the OSMLA case, the data are often contaminated with noises of various characteristics. They are also spontaneous passages spoken by speakers of various accents, with different types of recording equipment, etc. This variability and low quality imposes many problems for automatic speech labeling. Note, however, that all of this speech is easily understandable by human listeners.

HMM-based ASR performing forced alignment is the most popular method for phonetic labeling. Although the HMM method is not optimized for speech alignment, it is a well-studied technique and yields reasonably accurate alignment. Segmentation performance obtained with a HMM can be further improved using boundary refinement post-processing steps.

The objective of this work was to determine speech features, HMM parameters and HMM training method to align the OSMLA database at the phonetic level with highest accuracy. Optimum system parameters were first determined using the TIMIT database, and applied to the OSMLA database.

The remainder of this paper is organized as follow: Section 2 briefly describes the OSMLA database project. Section 3 discusses proposed methods, Section 4 explains experiments using TIMIT, Section 5 presents experiments on the OSMLA database, and Section 6 concludes the paper.

2. The open source multi-language audio database

The OSLMA database comprises a large database of English, Mandarin, and Russian audio/video recordings. The data were collected, formatted, organized, annotated, and given time aligned orthographic transcriptions at the sentence/phrase level by human listeners. Each language consists of 300 recordings, resulting in a total of approximately 90 hours of audio/video data. The speech data were often contaminated with noises of various characteristics, due to acoustic conditions of the recording site and the quality of the recording equipment. Moreover, the speakers speak with various accents and emotions. A detailed description of the database is presented in [1].

An ultimate goal of the database project is to obtain time-aligned transcriptions at the phonetic level of highest possible accuracy with the least human intervention. To help in this process, for each recording, 3 utterances (each 4-10 seconds long) were manually labeled at the phonetic level. This

information is meant to serve as ground truth test data for algorithm development, rather than for system training.

3. Proposed method

Two major difficulties that underlie this work are: quality of the speech, and unavailability of the correct phonetic sequences. As briefly discussed in Section 2 about the speech quality, a desirable automatic phonetic alignment system is one that is robust to noises and that can cope with dialectic variations.

The database was manually transcribed as word sequences with noise events marked. The only available time markers are at the sentence boundaries. Thus the only two inputs that were available to the automatic phonetic alignment system are the speech data and its sentence level transcription (sequence of words).

A hybrid architecture was proposed in [6] to automatically segment the TIMIT database using its word transcription (not using the phonetic transcription). The system was a HMM-based phone recognizer initialized with embedded parameter estimation. The initial models were used to force-align the training data. New HMMs were created and trained using the force-aligned phonetic labels for bootstrapping. The process was repeated for several passes to refine the HMM models. The alignment accuracy obtained with this method was 83.6%, using a 20 ms tolerance.

The proposed method uses HMM-based ASR running in forced alignment mode to achieve phonetic labeling. This labeling will serve as initial information for further boundary refinement. Our method is similar to [6] in that the input to the system is the speech signal and its corresponding word sequence, and the use of multi-pass training. Fig.1 illustrates the method.

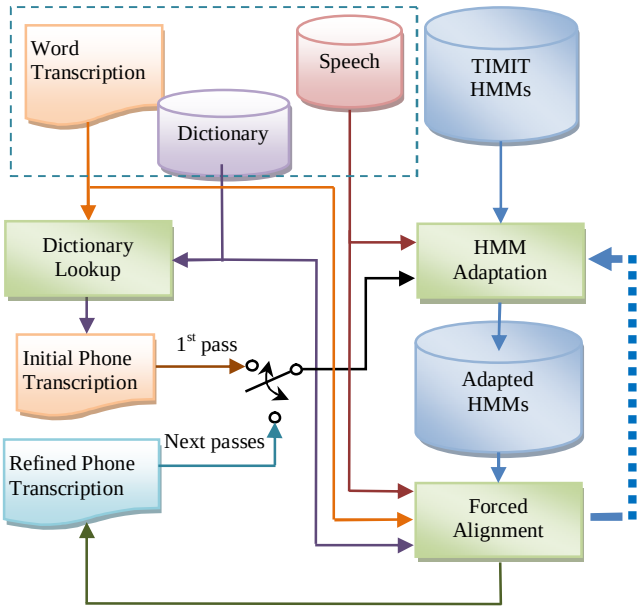

Figure 1: *The proposed method.*

In Figure 1, TIMIT HMMs represent phone-based HMMs trained with TIMIT data. In the first pass, the TIMIT models are adapted using the speech data and its word transcription, resulting in the adapted models (shown as Adapted HMMs in the figure). Since the adaptation needs phone transcriptions, the initial phone transcription is used as obtained from the word transcriptions by dictionary lookup (using the first pronunciation). The dictionary in Figure 1 refers to the pronunciation dictionary (lexicon) which maps each word to a phone sequence. The dictionary also supports words with multiple pronunciations.

The adapted models are used to perform forced alignment of the speech data using the word transcription and the dictionary. The outcome of this step is the refined phone transcription in which multiple pronunciation of words are taken into account. This is the by-product results of the Viterbi search in HMM decoding. Note that the dictionary used in this work was created such that for each pronunciation of a word, an extra phone sequence that ends with SP (short pause) was inserted. The purpose is to enable the forced alignment to automatically insert a short pause between a pair of words where there is high probability of an acoustic clue for a short pause.

In order to obtain better model parameters, the model adaptation process is repeated for several passes. For all but the initial pass, the refined phone transcriptions are used for model adaptation instead of the initial transcription.

Due to the noisy and dialectic nature of our database, there is no single phonetic alignment method that works well on all utterances. The proposed HMM-based ASR method yields an all around best result which would be valuable for further refinement. The lack of labeled data to train the system (used for most high-performance speech alignment systems) is not a severe limitation because of the availability of HMM adaptation techniques. So initial HMM models can be trained using labeled databases such as TIMIT.

Based on preliminary experiments, alignment accuracy obtained from adapted models was highest when adaptation is performed on a passage-by-passage basis. By passage, we mean the entire recording session (about 3-6 minutes long) of a speaker.

3.1. Figure of merit

In this work, two types of errors can occur: phonetic transcription errors and alignment errors. Both errors have to be accounted for at the same time. Alignment accuracy is measured along with phonetic transcription accuracy. Three type of transcription errors (insertion, deletion and substitution errors) are computed using a dynamic-programming-based string alignment algorithm. Only matched phones were used to compute alignment accuracy. The most commonly used figure of merit (as pointed out in [2]) for segmentation accuracy is the percentage of boundaries with errors smaller than 20 ms. This 20 ms error criteria was used in this work (although error tolerances of 5 ms, 10 ms, 30 ms, 50 ms and 100 ms were used in our experiments, results with 20 ms tolerance are mainly reported).

4. Experiments with TIMIT

The main purpose of this set of experiments was to investigate the best feature set, the best HMM parameter and the best HMM training that suits the task at hand. The TIMIT database was used in all initial experiments, as reported in this section. As mentioned above, since the proposed system is based on model adaptation from initial TIMIT models, it is important to determine a parameter set that works well with TIMIT.

According to the proposed system depicted in Figure 1, the system also behaves like a pattern recognizer when it is matching the acoustic information with the most likely phone sequence. Thus the adapted models must be a good recognizer. However some reported works have shown that the best

recognition architecture does not always result in best alignment accuracy, and vice versa [3, 10]. It would be beneficial to investigate the phone recognition and alignment performances simultaneously using different settings, to find the setting that gives balanced results between recognition accuracy and alignment accuracy.

4.1. Feature sets

The feature sets used in the experiments include: Mel-frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Perceptual Linear Prediction Coefficients (PLP), and Discrete Cosine Transform Coefficients (DCTC). For MFCC, LPCC and PLP, 13 coefficients were extracted from each frame and their corresponding delta and delta-delta terms were augmented resulting in 39 terms. These feature sets are referred to as MFCC_39, LPCC_39 and PLP_39. Note that all of these features were computed with 25 ms frame size with 5 ms frame spacing.

The DCTC features were computed as described in [11]. Two sets of DCTC features were investigated: DCTC_39 and DCTC_78. The DCTC_39 feature set was computed with 13 discrete cosine transform coefficients (DCTCs) of the log spectrum, with each DCTC trajectory encoded with 3 terms of a discrete cosine series (DCSs) over a block of 200 ms. The frame size was 25 ms and the frame space was 5 ms or 10 ms. The DCTC_78 features were computed with 13 DCTCs with 6 DCSs expansion, resulting in 78 terms. These features were computed using a frame size of 8 ms and expanded with DCSCs over a block of 200 ms with a block spacing of 8 ms. This DCTC_78 feature set was included in the experiments because it yielded excellent phone recognition accuracy on TIMIT test data.

4.2. HMM architecture

The HMM used in this set of experiment is *n*-state left-to-right model (no skipping), with *m* diagonal covariance Gaussian mixture components. The number of states, *n*, was 3, 4, 5 or 6 and the number of mixtures, *m*, was 1, 2, 4, 8, 16, or 32. The collapsed TIMIT's 48 phone set was used. The HTK toolkit [12] was used to build and train the models and also used to compute MFCC, LPCC and PLP features.

Two methods of HMM training were tested: isolated-word training and full training. The isolated-word training was achieved by using the isolated-word Viterbi training (available in HTK's HInit tool). No further training was performed. The full training utilized all HMM training steps i.e. isolated-word Viterbi training + isolated-word Baum-Welch training + embedded training. The isolated-word training method has been shown to be superior to full training for the phonetic alignment task [10].

4.3. Results

Tables 1, 2 and 3 show recognition results and alignment results with 8-mixture HMM models when the number of states is 3, 4 and 5, respectively, for the three tables. The alignment accuracies at 20 ms tolerances are reported.

Several observations can be drawn from Table 1, 2 and 3:

1. Feature sets that worked well for recognition also work well for alignment.
2. Alignment accuracies obtained with the isolated-word training were higher than for full training, while the recognitions accuracies were the other way around.
3. DCTC_78 yields all around best results with 4-state HMMs.

Table 1: *Results with 8-mixture models, 3 states.*

Feature set	% Accuracy			
	Isolated-word Trn.		Full Training	
	Recog.	Align	Recog.	Align
MFCC_39	62.9	86.9	65.5	85.0
LPCC_39	58.1	85.2	61.8	84.0
PLP_39	57.7	85.1	65.6	85.2
DCTC_39	65.6	87.6	66.9	81.8
DCTC_78	**67.1**	**89.6**	**68.2**	**85.6**

Table 2: *Results with 8-mixture models, 4 states.*

Feature set	% Accuracy			
	Isolated-word Trn.		Full Training	
	Recog.	Align	Recog.	Align
MFCC_39	63.8	87.9	66.7	86.6
LPCC_39	58.6	86.5	62.3	85.4
PLP_39	58.2	86.1	66.2	86.8
DCTC_39	65.9	88.5	67.4	83.6
DCTC_78	**66.7**	**90.4**	**68.5**	**87.6**

Table 3: *Results with 8-mixture models, 5 states.*

Feature set	% Accuracy			
	Isolated-word Trn.		Full Training	
	Recog.	Align	Recog.	Align
MFCC_39	63.7	88.8	66.4	**87.8**
LPCC_39	58.7	87.4	61.6	86.9
PLP_39	57.8	87.2	65.7	87.4
DCTC_39	**65.5**	**89.4**	**67.3**	84.9
DCTC_78	63.4	89.0	66.7	86.4

Observation 2 agrees with that reported in [10]. In addition, DCTC_78 features yielded even better results with more HMM mixture components (results not shown). Note that for all 39 feature parameter sets, results are only reported for a 5 ms frame spacing; results obtained with a 10 ms frame spacing were typically about 1-2% lower. Although DCTC_39 outperformed DCTC_78 with 5-state HMMS (Table 3), best absolute results were obtained with DCTC_78 and 4-state HMMs (Table 2).

Results in Table 4 indicate that all feature sets (except for the alignment accuracy of the MFCC_39) yield higher accuracy when trained with 32 mixture components. The DCTC_78 feature set yields the highest performance. Thus this feature set was chosen for further investigation in the next set of experiments.

Table 4: *Results with 32-mixture models, 4 states.*

Feature set	% Accuracy			
	Isolated-word Trn.		Full Training	
	Recog.	Align	Recog.	Align
MFCC_39	66.2	88.3	68.4	52.2
LPCC_39	60.3	86.6	64.3	85.9
PLP_39	58.9	86.1	67.5	86.3
DCTC_39	69.0	89.0	70.6	85.3
DCTC_78	**70.5**	**90.9**	**72.9**	**88.6**

5. Experiments using the open source database

The primary pronunciation dictionary used was CMU dictionary version 07a. This dictionary utilizes 39 phones; thus the 61

TIMIT phones were collapsed to 39. As a result, the TIMIT models consisted of 39 HMMs. Each HMM was modeled with 4 states and 32 mixtures, and trained with the full training method described in Section 4.2. Note that an additional dictionary was manually created to handle words that are not in the primary dictionary.

A subset of 212 passages (146 male + 66 female speakers) of the English speech from the OSMLA database was used. Each passage was first segmented into short utterances according to the utterance boundaries marked by human listeners. All noisy utterances were rejected. For each passage, there were 3 utterances that were manually labeled and they were used for performance evaluation.

The maximum likelihood linear regression (MLLR) HMM adaptation which is available in the HTK tool was used to adapt the TIMIT model using the speech data from each passage. The adaptation and evaluation was performed on a passage-by-passage basis. Sub-results were accumulated and reported as a whole after all passages had been processed.

The DCTC_78 and the MFCC_39 feature sets were tested. Table 5 shows all phone transcription error rates based on the total of 45,031 phones and Table 6 shows the alignment results based on the phone transcription errors in Table 5.

Table 5: *Phone transcription errors with MFCC_39 and DCTC_78 feature sets.*

Feature set	**Phone transcription error (%)**		
	Insertion	**Deletion**	**Substitution**
MFCC_39	1.59	**0.40**	1.63
DCTC_78	**0.64**	0.64	**1.20**

Table 6: *Alignment accuracies with MFCC_39 and DCTC_78 feature sets.*

Feature set	**Alignment Accuracy (%)**					
	5 ms	**10 ms**	**20 ms**	**30 ms**	**50 ms**	**100 ms**
MFCC_39	32.1	55.5	78.6	87.9	94.6	**98.4**
DCTC_78	**33.2**	**63.6**	**82.3**	**89.5**	**95.2**	98.2

From Table 5, DCTC_78 features yield lower insertion and substitution errors. In Table 6, DCTC_78 features yielded higher alignment accuracy than the MFCC-39 in all cases except for 100 ms tolerance.

6. Conclusions and future work

DCTC_78 features have been shown to be features of choice for ASR and speech alignment on the TIMIT and the OSMLA databases. The proposed HMM-based ASR forced alignment system that utilizes passage-by-passage adaptation is promising. Our future work will be focused on inclusion of noisy speech in the adaptation data set and aligning this speech. Robust boundary refine methods will be investigated to improve the alignment accuracy obtained from the HMM alignment.

6. Acknowledgement

This material is based on research sponsored by the Air Force Research Laboratory under agreement numbers FA8750-10-2-0160 and FA87501210093. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

7. References

[1] Zahorian, S. A. et al, "Open Source Multi-Language Audio Database for Spoken Language Processing Applications," Proc. INTERSPEECH-2011, pp.1493-1497, Florence, Italy, 2011.

[2] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L., TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993.

[3] Toledano, D. T. and Gomez, L. A. H., "Automatic phonetic segmentation," IEEE Trans. on Speech and Audio Processing, vol.11, no.6, pp.617-625, Nov., 2003.

[4] Hosom, J. P., "Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information," Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, 2000.

[5] Black, A. W. and Kominek, J., "Optimizing segment label boundaries for statistical speech synthesis," Proc. ICASSP, pp.3785-3788, 2009.

[6] Mporas, I., Ganchev, T. and Fakotakis, N., "A hybrid architecture for automatic segmentation of speech waveforms," Proc. ICASSP-2008, pp.4457-4460, Las Vegas, USA, 2008.

[7] Lin, C. Y., Jang, J. R. and Chen, K. T., "Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpora for Concatenation-based TTS," Computational Linguistics and Chinese Language Processing, Vol. 10, No. 2, pp. 145-166, June 2005.

[8] Jarifi, S., Pastor, D. and Rosec, O., "A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis," Speech Commun., vol.50, no.1, pp.67-80, 2008.

[9] Antonio, J. A. and Bonafonte, A., "Towards phone segmentation for concatenative speech synthesis,", Proc. 5th ISCA Speech Synthesis Workshop, pp139-144, 2004.

[10] Dines, J., Sridharan, S. and Moody, M., "Automatic speech segmentation with HMM," Proc. 9th Australian Int. Conf. on Speech Science & Technology, Melbourne, 2002.

[11] Zahorian, S. A., Hu, H., Chen, Z. and Wu, J., "Spectral and Temporal Modulation Features for Phonetic Recognition," Interspeech 2009, pp. 1071-1074, 2009.

[12] Young, S. J. et al, The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2006.