Automatic Phonetic Segmentation using Boundary Models

Jiahong Yuan[1], *Neville Ryant*[1], *Mark Liberman*[1]
Andreas Stolcke[2], *Vikramjit Mitra*[3], *Wen Wang*[3]
[1]University of Pennsylvania, [2]Microsoft Research, [3]SRI International

Abstract

This study attempts to improve automatic phonetic segmentation within the HMM framework. Experiments were conducted to investigate the use of phone boundary models, the use of precise phonetic segmentation for training HMMs, and the difference between context-dependent and context-independent phone models in terms of forced alignment performance. Results show that the combination of special one-state phone boundary models and monophone HMMs can significantly improve forced alignment accuracy. HMM-based forced alignment systems can also benefit from using precise phonetic segmentation for training HMMs. Context-dependent phone models are not better than context-independent models when combined with phone boundary models. The proposed system achieves 93.92% agreement (of phone boundaries) within 20 ms compared to manual segmentation on the TIMIT corpus. This is the best reported result on TIMIT to our knowledge.

Index Terms: phonetic segmentation, phone boundary model, forced alignment, HMM

1. Introduction

In the last twenty years, many large speech corpora have been collected for speech technology development. The ability to use speech corpora for research in other fields, such as phonetics, sociolinguistics and psychology, depends on the availability of phonetic segmentation and transcriptions. Manual phonetic segmentation is time-consuming and expensive; it could take as long as 400 times real time [1] or 30 seconds per phone [2]. Automatic phonetic segmentation is much needed.

A widely used method for automatic phonetic segmentation is "forced alignment". The task requires two inputs: recorded audio and phone or word transcriptions. If only word transcriptions are available, the transcribed words are mapped into a phone sequence in advance by using a pronouncing dictionary, or grapheme to phoneme rules. The most common approach for forced alignment is to build a Hidden Markov Model (HMM) based phonetic recognizer [3-8]. In this approach, each phone is a HMM that has typically 3-5 states. The speech signal is analyzed as a successive set of frames. The alignment of frames with phones is determined by finding the most likely sequence of hidden states (which are constrained by the known sequence of phones) given the observed data and the acoustic model represented by the HMMs. The reported performances of standard HMM-based forced alignment systems range from 80%-89% agreement (of all boundaries) within 20 ms compared to manual segmentation on the TIMIT corpus [6].

A main drawback of the HMM-based forced alignment for phonetic segmentation is that phone boundaries are not represented in the model. The boundaries are simply derived from the alignment of phone states with frames. This is different from the manual phonetic segmentation process, in which the acoustic landmarks at phone boundaries [9], e.g., an abrupt spectral change, are used to determine the location of a boundary. Many researchers have tried to overcome this drawback. One method is to take a two-stage scheme, where HMM-based forced alignment is followed by local boundary refinement. For example, [10] used energy changes in different frequency bands for boundary correction, [11] trained support vector machine (SVM) classifiers to differentiate boundaries from non-boundary positions, and [12, 13] employed neural network to refine phone boundaries. [14] described a non-HMM system for phone alignment based on discriminative learning. In their system a set of base functions were learned to measure the confidence for an alignment. [15] proposed several modifications to an HMM-based system, including the use of energy-based features and distinctive phonetic features, and the use of observation-dependent state transition probabilities. The proposed system of [15] achieved 93.36% agreement within 20 ms compared to manual segmentation, which is the best known reported result on the TIMIT corpus. Human labelers have an average agreement of 93% within 20 ms on various corpora, and an agreement of 93.49% within 20ms on TIMIT [6, 15].

In this study, we investigate the use of phone boundary models for forced alignment within the HMM framework. The idea is to treat phones and phone boundaries as independent HMMs. A boundary is determined by the alignment of its own state with frames. Three related questions are also investigated: (i) whether to use context-independent (monophone) or context-dependent (triphone) phone models in terms of forced alignment performance; (ii) is it helpful to use precise phonetic segmentation for training HMMs? (iii) is it helpful to separate cross-word and within-word phone boundaries? In forced alignment, unlike in automatic speech recognition, monophone HMMs are more commonly used than triphone HMMs. [7] showed that monophone models outperform triphone models in forced alignment for medium tolerances (15-30 ms different from manual segmentation), while underperform for small (5-10 ms) and large tolerances (>35 ms). A possible reason is that a context-dependent HMM does not have information to discriminate between the target phone and its context; therefore part of other phones may be modeled by the HMM [4, 7]. In this case, phone boundary accuracy will be sacrificed although phone recognition accuracy may not. To force context-dependent HMMs not to model part of other phones, we can use the acoustic observations of individual phones, instead of entire utterances with transcribed phone sequences, as input for training HMMs. In HMM-based speech recognition manual phonetic segmentation is generally not used; forced alignment segmentation seems to be precise enough for training HMMs because HMM training is an averaging process that tends to smooth segmentation errors [7]. The goal of automatic phonetic segmentation is, however, different from that of speech recognition. It remains unknown whether the use of precise phonetic segmentation for training HMMs can improve the performance of forced alignment. Finally, many studies have demonstrated that there are acoustic cues (though not

infallible) to word boundaries [16-18] and the cues are used for word segmentation by listeners [19-21]. We investigate in this study whether within-word and cross-word boundaries should be separated in phone boundary models used for forced alignment.

In the following sections we first introduce the data set and the evaluation method in Section 2. In Section 3 we present experiments that address the questions posed above. Our proposed system is described in Section 4, followed by a discussion of future research in Section 5.

2. Data and Evaluation

The TIMIT corpus was used [22]. Excluding the "dialect calibration" sentences (SA sentences), 3,696 utterances from the training partition of the corpus were used for training and 1,344 utterances from the test partition were used for testing. Following [15], the 61 TIMIT phonemes were mapped to 54 phonemes (detailed description on page 357 of [15]). The syllabic phonemes /em/, /en/, /eng/, and /el/ were mapped to their non-syllabic counterparts /m/, /n/, /ng/, and /l/. The glottal closure symbol /q/ was removed. It was either merged with the neighboring voiced phonemes or replaced with a schwa /ax/ if surrounded by two unvoiced phonemes. Short pauses with duration less than 20 ms were also removed and merged with neighboring phonemes. The 54 phonemes are listed in Table 1. The boundaries between two pauses, including stop closures, were excluded from evaluation. There were 136,450 boundaries in the training set, and 49,248 boundaries in the test set for evaluation.

Table 1. *The phoneme set (54 phonemes).*

Pauses and stop closures	/pau/, /pcl/, /bcl/, /tcl/, /dcl/, /kcl/, /gcl/
Vowels	/aa/, /ae/, /ah/, /ao/, /aw/, /ax/, /axh/, /axr/, /ay/, /eh/, /er/, /ey/, /ih/, /ix/, /iy/, /ow/, /oy/, /uh/, /uw/, /ux/
Glides	/l/, /r/, /w/, /y/, /hh/, /hv/
Nasals	/m/, /n/, /ng/, /nx/
Plosives	/b/, /d/, /g/, /p/, /t/, /k/, /dx/, /jh/, /ch/
Fricatives	/s/, /z/, /sh/, /zh/, /f/, /v/, /th/, /dh/

In our experiments, forced alignment boundaries were adjusted by two statistical correction procedures before evaluation, one for the boundaries between two phonemes of vowels or glides, and one for the other boundaries. The boundaries between vowel/glide phonemes are inherently subjective. The criteria for TIMIT boundary assignments stated that ([23]): "The boundary between many semivowels and their adjacent vowels is rather ill-defined in the waveform and spectrogram, because transitions are slow and continuous. It is not possible to define a single point in time that separates the vowel from the semivowel. In such case we decided to adopt a simple heuristic rule, in which one-third of the vocalic region is assigned to the semivowel". To compensate for such arbitrariness, we built a linear model to correct the forced-alignment boundaries between vowel/glide phonemes. The model predicts manual boundary positions from the forced alignment positions of the two phonemes (phoneme center positions), the identities of the boundaries (the phonemes preceding and following the boundary), and the forced alignment boundary positions. The model was trained on the training data and applied to the test data. For all other boundaries, the mean difference between manually labeled and forced alignment boundaries for every boundary identity was calculated using the training data, and the forced alignment boundaries in the test set were shifted by these boundary-dependent time differences. This correction is to compensate for the systematic segmentation errors produced by HMMs.

The accuracy of automatic segmentation is generally measured in terms of what percentage of the automatically labeled boundaries are within a given time threshold (tolerance) of the manually labeled boundaries. 20 ms has been most widely used as a tolerance for measuring phone segmentation quality. In the following experiments the agreement percentages for 10 to 50 ms tolerances are reported.

3. Experiments

3.1. Precise phone segmentation for training

To utilize manual phone segmentation for training HMMs, we obtained the acoustic observations of individual phones by extracting frames within the phone boundaries from observations (features) of utterances. The observations of individual phones were then used for training. As a comparison, the observations of entire utterances and their phone transcriptions were also used for training (which is a common practice in speech recognition). Monophone HMM and GMM acoustic models, with the standard 39 PLP features extracted with 25ms Hamming window and 10ms frame rate [24], were trained using the HTK toolkit [25]. The number of states in the HMM models and the number of Gaussian mixtures were optimized for best alignment performance with 20ms tolerance. Stops, stop closures, the vowel /axh/ ("devoiced schwa"), nasals, and liquids (/l/, /r/) are 1-state HMMs; the "true" diphthongs (/ay/, /aw/, /oy/) are 5-state HMMs; and the other phonemes are 3-state HMMs. Eight Gaussian mixtures were used. In testing, forced alignment was run over utterances given their phone transcriptions. The forced alignment boundaries were adjusted by applying the statistical correction procedures described in Section 2. The results, in terms of agreement between forced alignment and manually labeled boundaries, are listed in Table 2.

Compared to the system trained on utterances, the system trained on individual phones increased forced alignment accuracy by 3.03% for 10ms tolerance (from 70.20% to 73.23%), by 1.87% for 20ms tolerance (from 89.98% to 91.85%), and by 0.13% for 50ms tolerance (from 98.92% to 99.05%). The relative error reductions for 10ms, 20ms, and 50ms tolerances are 10.2%, 18.7%, and 12.0% respectively. Clearly, using precise phonetic segmentation for training HMMs can significantly improve forced alignment quality. In the following experiments phone HMMs are trained on the observations of individual phones.

Table 2. *Agreement percentages for different tolerances (in ms), for systems using or not using manual segmentation for training monophone HMMs.*

	<10	<20	<30	<40	<50
Segmentation not used for training	70.20	89.98	95.74	97.88	98.92
Segmentation used for training	**73.23**	**91.85**	**96.45**	**98.17**	**99.05**

3.2. Context dependent and independent models

A within-word triphone model was used to compare with the context-independent monophone model. The triphone HMMs had the same number of states as their monophone counterparts. The triphone states were tied using decision trees, and the degree of state-tying was optimized for best alignment performance with 20ms tolerance. 792 tied states were used.

The results of the triphone model with 1, 2, 4, and 8 Gaussian mixtures are listed in Table 3, for comparison with the monophone model with 8 Gaussian mixtures. From the table we can see that, on one hand, the triphone model outperforms the monphone model for all tolerances. This result is different from [7], which found that monophone models outperform triphone models for medium tolerances (15-30ms). This may be due to the fact that in our experiment HMMs were trained on individual phones not utterances. On the other hand, however, the differences between the triphone and monophone models were relatively small. The absolute error reductions for 10ms, 20ms, and 50ms tolerances are 1.86%, 0.52%, and 0.1% respectively; and the relative error reductions for these tolerances are 6.95%, 6.38%, and 10.5% respectively. The results in Table 3 also suggest that fewer Gaussian mixtures are beneficial for medium or small tolerances while more Gaussian mixtures are beneficial for large tolerances. This result is consistent with [7].

Table 3. *Agreement percentages for different tolerances (in ms), for systems using monophone HMMs with 8 Gaussian mixtures and using triphone HMMs with 1, 2, 4, and 8 Gaussian mixtures.*

	<10	<20	<30	<40	<50
Monophones	73.23	91.85	96.45	98.17	99.05
Triphones (1GMM)	74.53	92.31	96.63	98.34	99.12
Triphones (2GMM)	74.93	**92.37**	96.72	98.33	99.09
Triphones (4GMM)	**75.09**	92.17	96.41	98.13	98.92
Triphones (8GMM)	73.61	92.17	**96.76**	**98.40**	**99.15**

3.3. Phone boundary models

In addition to phone HMMs, phone boundary HMMs were also trained in this experiment. Three types of boundary HMMs were tried: a 3-state HMM, a 1-state HMM, and a special 1-state model. The first two are typical HMMs whereas the last one is not a true Markov chain. Figure 1 illustrates a typical 1-state HMM, in which the transition probabilities $a_{01} = 1$ and $0 < a_{11}, a_{12} < 1$. The state can either repeat itself or exit from the model. In the special 1-state model for the boundaries, however, $a_{01} = 1$, $a_{11} = 0$ and $a_{12} = 1$. Therefore, a boundary can have one and only one state occurrence, i.e., aligned with only one frame. Because boundaries are not independent units and do not have time spans in TIMIT, we cannot use manually labeled individual boundaries for training boundary HMMs. To select the best phone boundary model for forced alignment, we trained boundary HMMs, together with monophone HMMs, on utterances. The results showed that the special 1-state model had the best alignment performance with 20ms tolerance.

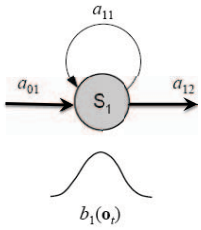

Figure 1: *1-state HMM. The one state HMM is a special 1-state model for the boundaries when the transition probabilities $a_{11} = 0$ and $a_{12} = 1$.*

Using the special 1-state model, we rebuilt phone boundary models by using the frames extracted at the boundaries, one frame for each boundary. Within-word and cross-word boundaries were differentiated. The full set of boundary models contained 5,832 states, one state for each boundary type (54 phonemes on the left * 54 phonemes on the right * 2). The boundary states were tied using decision trees, and the degree of state-tying was optimized for best alignment performance with 20ms tolerance. 734 tied states were used. To combine with the boundary models for forced alignment, monophone and triphone models were also retrained by excluding boundary frames from phone boundaries. In testing, forced alignment was run over utterances given their phone transcriptions and the boundaries between phones. The forced alignment boundaries were adjusted by applying the same statistical correction procedures.

The results of using phone boundary models are listed in Table 4. Compared to the system using monophone HMMs, the system using both monophone HMMs and boundary models increased forced alignment accuracy by 4.21% for 10ms tolerance (from 73.23% to 77.44%), by 2.07% for 20ms tolerance (from 91.85% to 93.92%), and by 0.3% for 50ms tolerance (from 99.05% to 99.35%). The relative error reductions for 10ms, 20ms, and 50ms tolerances are 15.7%, 25.4%, and 31.6% respectively. The system using both triphone HMMs and boundary models, compared to the system using triphone HMMs only, increased forced alignment accuracy by 3.16% for 10ms, 1.48% for 20ms, and 0.28% for 50ms. The relative error reductions are 12.6%, 19.4%, and 30.8%. We note that although triphone HMMs slightly outperform monophone HMMs for all tolerances (shown in Table 3), the combination of monophone HMMs and boundary models outperforms the combination of triphone HMMs and boundary models for 20-40 ms tolerances.

Table 4. *Agreement percentages for different tolerances (in ms), for systems using monophone HMMs, monophone HMMs and boundary models, triphone HMMs, and triphone HMMs and boundary models.*

	<10	<20	<30	<40	<50
Monophones	73.23	91.85	96.45	98.17	99.05
Monophones & Bo.	77.44	**93.92**	**97.43**	**98.78**	99.35
Triphones	74.93	92.37	96.72	98.33	99.09
Triphones & Bo.	**78.09**	93.85	97.37	98.72	**99.37**

Finally, we built a new set of boundary models in which within- and cross- word boundaries were not differentiated. The results are listed in Table 5. The system differentiating

within- and cross- word boundaries had slightly better performance except for the 10ms tolerance. However, the differences between the two systems are very small.

Table 5. *Agreement percentages for different tolerances (in ms), for systems differentiating and not differentiating within- and cross- word boundaries.*

	<10	<20	<30	<40	<50
Word boundaries differentiated	77.44	**93.92**	**97.43**	**98.78**	**99.35**
Word boundaries not differentiated	**77.53**	93.84	97.39	98.75	99.35

4. The proposed system

From the experiments above, we can conclude that the use of explicit phone boundary models within the HMM framework can significantly improve forced alignment accuracy. HMM-based forced alignment systems can also benefit from using precise phonetic segmentation for training HMMs. Figure 2 shows the two improvements, compared to the baseline system in which monophone HMMs were trained on utterances.

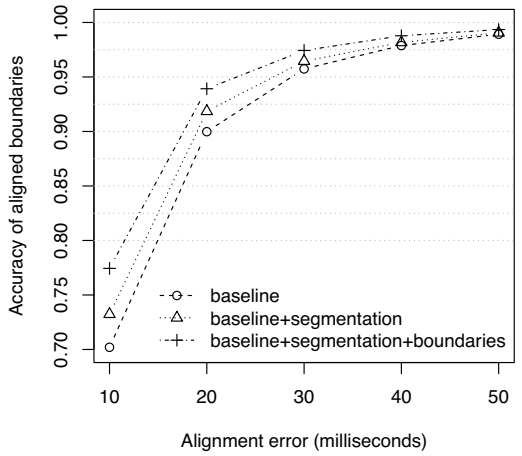

Figure 2: *The improvements over the baseline system by using precise phone segmentation for training HMMs and by using phone boundary models.*

The proposed system is summarized in Table 7. The system achieves 93.92% agreement within 20ms. It improves from the baseline system by 3.94%, and from the best reported result in the literature (93.36%) by 0.56%. The relative error reductions are 39% over the baseline system and 8% over the best reported result.

The most frequent errors in the results of the proposed system are boundaries between vowel/glide phonemes, and sentence-final boundaries, especially when the last phone is a nasal. The boundaries in the test set that have 30 or more errors for 20ms tolerance are listed below:

/n/_/pau/: 55	/aa/_/r/:51	/ao/_/r/: 47
/pau/_/dh/: 45	/l/_/iy/: 43	/ao/_/l/: 39
/m/_/pau/: 37	/iy/_/pau/: 34	/eh/_/r/: 30

Two authors of the paper, MYL and NR, manually labeled 100 offset boundaries of sentence-final nasals that were randomly selected from the test set. The agreements within 20 ms between the two authors, between the authors and TIMIT, and between the authors and the proposed system are all relatively low and at the same level, as shown in Table 6. This result suggests that more consistent phonetic segmentation strategies should be adopted for constructing a database that can be used to evaluate further improved automatic phonetic segmentation techniques.

Table 6. *Agreements within 20ms between two authors, TIMT and the proposed system on 100 offset boundaries of sentence-final nasals.*

	MYL	NR	TIMIT	System
MYL	-	51	47	46
NR	51	-	50	57
TIMIT	47	50	-	54
System	46	57	54	-

Table 7. The proposed system.

Phone models: Monophone HMMs.
1-state HMMs: /pcl/, /bcl/, /tcl/, /dcl/, /kcl/, /gcl/, /axh/, /l/, /r/;
5-state HMMs: /ay/, /aw/, /oy/;
3-state HMMs: all other phonemes;
Trained on individual phones.
Boundary models: Special 1-state model in which the transition probability for the state repeating itself is 0.
Differentiating within- and cross- word boundaries;
Boundary states tied to 734 states;
Trained on boundary frames, one frame for each boundary.
Statistical correction to forced alignment boundaries.
Boundaries between vowel/glide phonemes are corrected by a linear model;
Other boundaries are shifted by a boundary-dependent time difference;
Both are trained on training data.

5. Future research

Unlike TIMIT, most large speech corpora don't have phonetic transcriptions. It is a more challenging task for forced alignment when only word transcriptions are available. Natural speech is highly variable, simple word-to-phoneme mapping (either by using a pronouncing dictionary or grapheme to phoneme rules) may not always generate phone sequences that contain the correct pronunciation. Moreover, transcribing words in spontaneous speech is itself a difficult task. Disfluencies, for example, are often missed in the transcription process. Future research needs to address the issues of pronunciation variation (e.g., deletion, reduction, and insertion), disfluencies and imperfect transcription in automatic phonetic segmentation.

6. Acknowledgements

This work is supported in part by NSF grant IIS-0964556.

7. References

[1] Godfrey, J. J., Holliman, E. C. and McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development," *Proceedings of ICASSP 1992*, pp. 517-520. Revision, February 19, 1997.

[2] Leung, H. and Zue, V.W., "A procedure for automatic alignment of phonetic transcription with continuous speech," *Proceedings of ICASSP 1984*, pp. 73-76, 1984.

[3] Brugnara, F., Falavigna, D. and Omologo, M., "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication,* 12, pp. 357–370, 1993.

[4] Ljolje, A., Hirschberg, J. and van Santen, J, "Automatic speech segmentation for concatenative inventory selection," in J. van Santen, R. Sproat, J. Olive and J. Hirschberg (ed.), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 305-311, 1997.

[5] Wightman, C. and Talkin, D., "The Aligner: Text to speech alignment using Markov Models," in J. van Santen, R. Sproat, J. Olive and J. Hirschberg (ed.), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 313-323, 1997.

[6] Hosom, J.P., *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. PhD thesis, Oregon Graduate Institute of Science and Technology, 2000.

[7] Toledano, D.T., Gomez, L.A.H. and Grande, L.V., "Automatic phoneme segmentation," *IEEE Trans. Speech and Audio Proc.*, 11, pp. 617-625, 2003.

[8] Yuan, J. and Liberman, M., "Speaker identification on the SCOTUS corpus," *Proceedings of Acoustics 2008*, pp. 5687-5690, 2008.

[9] Stevens, K., "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, 111, pp. 1872-1891, 2002.

[10] Kim, Y.-J. and Conkie, A., "Automatic segmentation combining an HMM-based approach and spectral boundary correction," *Proceedings of ICSLP 2002*, pp.145-148, 2002.

[11] Lo, H.-Y. and Wang, H-M., "Phonetic boundary refinement using support vector machine," *Proceedings of ICASSP 2007*, pp. 933-936, 2007.

[12] Toledano, D.T., "Neural network boundary refining for automatic speech segmentation," *Proceedings of ICASSP 2000*, pp.3438-3441, 2000.

[13] Lee, K.-S., "MLP-based phone boundary refining for a TTS database," *IEEE Trans. Audio, Speech, and Language Proc.*, 14, pp. 981-989, 2006.

[14] Keshet, J., Shalev-Shwartz, S., Singer, Y. and Chazan, D., "Phoneme alignment based on discriminative learning." *Proceedings of Interspeech 2005*, pp. 2961–2964, 2005.

[15] Hosom, J.P., "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, 51, pp. 352-368, 2009.

[16] Lehiste, I., "An acoustic-phonetic study of internal open juncture," *Phonetica*, 5, supplement, pp. 5-54, 1960.

[17] Turk, A.E. and Shattuck-Hufnagel, S., "Word-boundary-related duration patterns in English," *Journal of Phonetics*, 28, pp. 397-440, 2000.

[18] Garellek, M., "Word-initial glottalization and voice quality strengthening," *UCLA Working Papers in Phonetics*, 111, pp. 92-122, 2012.

[19] Nakatani, L.H. and Dukes, K.D., "Locus of segmental cues for word juncture," *J. Acoust. Soc. Am.*, 62, pp. 714-719, 1977.

[20] Johnson, E.K. and Jusczyk, P.W., "Word segmentation by 8-month-olds: when speech cues count more than statistics," *Journal of Memory and Language*, 44, pp. 548-567, 2001.

[21] Tyler, M.D. and Cutler, A., "Cross-language differences in cue use for speech segmentation," *J. Acoust. Soc. Am.*, 126, pp. 367-376, 2009.

[22] Garofolo, J.S., *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (LDC93S1), Linguistic Data Consortium, 1993.

[23] Zue, V.W. and Seneff, S., "Transcription and alignment of the TIMIT database," in H. Fujisaki (ed.), *Recent Research Towards Advances Man-Machine Interface*, pp. 515-525, 1996.

[24] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech*," J. Acoust. Soc. Am.*, 87, pp. 1738-1752, 1990.

[25] The Hidden Markov Model Toolkit (HTK): http://htk.eng.cam.ac.uk/.