Automatic Phonetic Segmentation and Pronunciation Detection with Various Approaches of Acoustic Modeling

Petr Mizera and Petr Pollak(✉)

Faculty of Electrical Engineering, Czech Technical University in Prague,
K13131, Technicka 2, 166 27 Praha 6, Czech Republic
{mizera,pollak}@fel.cvut.cz
www.fel.cvut.cz
www.noel.feld.cvut.cz/speechlab

Abstract. The paper describes HMM-based phonetic segmentation realized by KALDI toolkit with the focus on study of accuracy of various acoustic modeling such as GMM-HMM vs. DNN-HMM, monophone vs. triphone, speaker independent vs. speaker dependent. The analysis was performed using TIMIT database and it proved the contribution of advanced acoustic modeling for the choice of a proper pronunciation variant. For this purpose, the lexicon covering the pronunciation variability among TIMIT speakers was created on the basis of phonetic transcriptions available in TIMIT corpus. When the proper sequence of phones is recognized by DNN-HMM system, more precise boundary placement can be then obtained using basic monophone acoustic models.

Keywords: Automatic phonetic segmentation
Pronunciation variability · GMM-HMM · DNN-HMM · KALDI
TIMIT

1 Introduction

Automatic phonetic segmentation is a procedure which defines boundary locations of particular phones in a given utterance and whose usage is necessary in situations when phone boundaries must be found for very huge corpora. It is typically used for a creation of subword units for the purpose of concatenative speech synthesis [8,13], for a determination of phone boundaries in huge speech corpora for the training of neural-networks-based speech recognition systems, or in other applications motivated by a study of pronunciation variability based on the analysis of phonetic segmentation results. Detailed analysis of particular phone realizations can also contribute to the clinical diagnostics of serious diseases which influence speech production, or to an analysis of pronunciation variability in spontaneous or informal speech [9].

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 419–429, 2018.
https://doi.org/10.1007/978-3-319-99579-3_44

The basic solution applied to the determination of phone boundaries is based on forced-alignment of trained HMM models for a given utterance with available acoustic realization and known content, optimally, at phonetic level. This procedure is standardly used as a significant step during the training of acoustic models of speech recognizers. It can be realized by various toolkits which implement HMM-based speech recognition, e.g. HTK [17], Sphinx [1], RWTH [14], or KALDI [12]. Especially KALDI is nowadays one of the most popular toolkits used world-wide by the speech research community.

In this paper, we present the analysis of phonetic segmentation accuracy using KALDI toolkit. We use acoustic models available in the standard KALDI TIMIT recipe, however, we work with more common setup when the phonetic content is not known. Many previously published approaches based on TIMIT corpus worked with available phone boundaries and many of them used known phonetic content for each utterance as the input of forced-alignment. Finally, we analyzed the accuracy of boundary determination as well as the precision of the choice of proper pronunciation variant when the transcription is available at word level and higher pronunciation variability is supposed in realized utterances.

2 Method

As was mentioned above, KALDI toolkit is frequently used for speech recognition by research community and consequently it is under continuous development. Currently, it covers many contemporary advanced techniques used within particular modules of ASR, including advanced techniques of acoustic modeling, mainly DNN-based systems. However, the usage of KALDI for speech segmentation is not so frequent [7]. Mentioned availability of advanced acoustic modeling techniques in KALDI was the main reason for this study describing an analysis how they benefit the precision of phonetic segmentation.

2.1 Phonetic Segmentation

Concerning the boundary determination, we used rather standard approach of forced-alignment. Its implementation using KALDI toolkit allowed us to study techniques using various approaches to acoustic modeling used in typical solutions ("recipes") available within KALDI distribution. Generally, we selected AM models which were suitable for generating targets for DNN-HMM training. We experimented with often used GMM-HMM models [12], i.e. the basic and simplest AM based on monophones (marked in the following text by acronym *mono*), speaker-independent triphone AM using basic short-time cepstral features (acronym *tri1*), speaker-independent triphone model with LDA features (acronym *tri2*), speaker-dependent triphone AM obtained by fMLLR and speaker-adaptive training (acronym *tri3*). In the end, the most advanced AM used in this work was DNN-HMM model (acronym *dnn*) with the topology of neural network consisting of the input layer with 440 units followed by 6 hidden layers with 2048 neurons per layer. The process of building of DNN-HMM

system started with the initialization of hidden layers by Restricted Boltzmann Machines and it was closed by frame cross-entropy training [4].

More advanced AM models based on time-delay neural networks with the lattice-free version of the maximum mutual information or long-short-term-memory networks [11] were not experimented. They help typically for an improvement of WER in speech recognition, but they are not so good for determination of phone boundaries using forced alignment[1].

Speech features were computed in accordance to the setup used in KALDI recipes. As basic cepstral features (used in AMs *mono* and *tri1*), we used 13 mel-frequency cepstral coefficients including zeroth cepstral coefficient, computed for the short-time frame with the length of 25 ms and shifted over the signal with the step of 10 ms. Cepstral-mean normalization was applied to this 13-element vector of static short-time features and they were completed by delta (dynamic) and delta-delta (acceleration) features to the final length of 39. LDA features (used in AM *tri2*) were computed from the context obtained by splicing of 5 short-time-feature vectors to both sides and followed by LDA and MLLT realizing decorrelation and the reduction of dimension to the length of 40. For AM *tri3*, it was followed by feature-space maximum likelihood linear regression (fMLLR) per each speaker (also called speaker adaptive training, SAT). Finally, these 40 dimensional fMLLR features with mean and variance normalization extended in both-side context were used as the input in *dnn* AM.

2.2 Impact of Pronunciation Lexicon

The accuracy of forced-alignment technique used for phonetic segmentation relies on the quality of inputs. Of course, it means the quality of acoustic data, however, it also depends strongly on the accuracy of input phonetic contents. Phonetic content of utterances transcribed usually at orthographical level can be obtained by grapheme-to-phoneme conversion or from a pronunciation lexicon, which can cover also a pronunciation variability [2] by including more pronunciation variants. This approach must be definitely used when phone boundaries should be determined for spontaneous and informal speech, higher diversity of language dialects, as well as in other situations when the level of pronunciation variability is rather high [18]. It can be obtained manually (for some very specific situations) or automatically (to extend regular pronunciations by particular phone substitutions or reductions on the basis of defined rules [9,10]).

In presented work, we analyzed the accuracy of phone boundaries determination in the case when the lexicon contained more pronunciation variants. For this purpose, we have created the lexicon containing all pronunciations which had appeared within phonetic transcription of TIMIT corpus (called further as *timit-variants*). It was obtained from available transcription at the word and phone level, i.e. as the new word pronunciation we took the sequence of all phones which lied within the word boundaries. Finally, the significant majority

[1] Discussed in KALDI community at https://groups.google.com/forum/#!topic/kaldi-help/cSAm5iXGhZo.

of words from TIMIT had more than one pronunciation, so we could analyze also the ability of used AM to recognize the correct pronunciation variant for particular word realizations. In total, we obtained 19184 pronunciations for 6256 words, moreover, in some cases the number of pronunciation variants was very high (22 words have more than 20 pronunciations), as it is shown in more details in Table 1. This lexicon should simulate using TIMIT corpus a realistic situation of phonetic segmentation of informal speech when each word can have more pronunciations due to pronunciation variability in informal speaking style.

Table 1. Lexicon timit-variants - statistics.

No. of pronunciation variants	1	2	3–5	6–10	11–20	21–50	65
No. of words	631	3372	1516	637	78	21	1

When pronunciation lexica contain such a high number of pronunciation variants (20 and more), correct detection of the proper pronunciation variant is very important task and phonetic segmentation in this setup can also serve to detect proper pronunciation variants within an analyzed utterance. It can then play the important role in the research focused on pronunciation variability and it was also analyzed in this work.

3 Experiments

The experimental part of this research was focused on the analysis of phonetic segmentation accuracy from the following three aspects: the optimum choice of proper acoustic model, the impact of extended pronunciation lexicon, and finally, the accuracy of pronunciation variant detection when more variants are available in the lexicon.

3.1 Used Tools and Speech Databases

All experiments were realized on the basis of TIMIT corpus [3], used often as a standard for the evaluation of phoneme classification, phoneme recognition, or phonetic segmentation for English. As it was mentioned above, designed acoustic model systems were built using KALDI toolkit.

Table 2. TIMIT data sets used in presented evaluations.

Data set	Speakers	Sentences	Hours	Num. words	Num. boundaries
TRAIN	462	3696	3.14	30132	-
CORE test set	24	192	0.16	1570	7215
COMPLETE test set	168	1344	0.81	11025	50754

We started with a standard s5 recipe for TIMIT available in KALDI distribution and we optimized it with regard to improve the accuracy of automatic phonetic segmentation task. The published recipe has been designed mainly for phoneme recognition task and it works with reference train and CORE test sets. For the phonetic segmentation task, we generated TIMIT COMPLETE test set with 168 speakers and 1344 test sentences. The phonetically-compact sentences (marked as SX sentences) and phonetically-diverse ones (marked as SI sentences) were only used for our experiments. TIMIT phoneme set was reduced from 61 to 48 final phonemes, which were used for acoustic modeling. The reduction to 39 phones was used finally for boundaries scoring as it is used standardly for English in KALDI recipes as well as by many other authors in ASR systems [6]. HMM topology consisted from 3 emitting states models for non-silence phonemes and 5 emitting states models for silence and direct phoneme transcription, which included also silence marks, was used for training AMs. Therefore silence appeared in training graphs and silence boundaries were scored, the optional silence was not used for our experiments. Finally, we used 50754 boundaries from COMPLETE test set and 7215 boundaries from CORE test set for our evaluations. The summary of used data sets is presented in Table 2.

3.2 Evaluation Criteria

The evaluation of phonetic segmentation accuracy was done using the criteria describing both the accuracy at the level of phone recognition correctness as well as the accuracy of phone boundary placement (as it was similarly used by other authors, e.g. [5,7]).

First, the phone recognition correctness is evaluated standardly using *Phone Error Rate* computed on the basis of Levenshtein distance as

$$PER = \frac{S + D + I}{N} \cdot 100 \tag{1}$$

where N is the number of phones in the reference and S, D, and I are the numbers of substitutions, deletions, and insertions in aligned data. It is also suitable to evaluate *Phone Correctness* computed as

$$PCorr = \frac{N - S - D}{N} \cdot 100 \tag{2}$$

because the evaluation of the accuracy of particular boundary placement makes sense just for correctly recognized phones. For further evaluations, all deleted phones are removed from the reference transcript, inserted phones from aligned transcript, and substituted phones are removed from both of them.

The cleared transcripts are then used for the evaluation of boundary placement accuracy. When we have two pairs of reference and transcribed boundaries for each phone realization, i.e. $beg_{ph,ref}[i]$ and $end_{ph,ref}[i]$ vs. $beg_{ph}[i]$ and $end_{ph}[i]$, the following two criteria *Phone Beginning Error (PBE)* and *Phone End Error (PEE)* can be defined as

$$PBE_{ph}[i] = |\, beg_{ph}[i] - beg_{ph,ref}[i] \,|\,, \tag{3}$$

$$PEE_{ph}[i] = |\, end_{ph}[i] - end_{ph,ref}[i] \,|\,. \tag{4}$$

The accuracy of phone boundary can be approximated using the rate of phone boundary error which is below the chosen threshold which can be defined as

$$PBE_{ph,thr} = \frac{\sum_{i=1}^{N_{ph}} (PBE_{ph}[i] < thr)}{N_{ph}} \tag{5}$$

where ph is phone/class identification, N_{ph} is the number of phone/class realizations, and thr is the value of chosen error threshold. Similarly, same procedure is applied for the computation of $PEE_{ph,thr}$. Threshold values used for realized evaluations within this work were 5, 10, 20, or 30 ms respectively.

All of these criteria can be computed with basic statistics for all particular phones, however, more often is the usage of their evaluation over defined phone classes, which are generally language independent. We used phone classes for English according to [5], i.e. VOW - vowels, GLI - semivowels and glides, VFR - voiced fricatives, UFR - unvoiced fricatives, NAS - nasals, STP - stops, UST - unvoiced stops, and SIL - silence.

Finally, we define *PronER* (Pronunciation Error rate) to evaluate pronunciation detection accuracy

$$PronER = \frac{S}{N} \cdot 100 \tag{6}$$

where N is the total number of words in the reference set and S is the numbers of incorrectly recognized (substituted) pronunciation variants.

3.3 Results

3.3.1 Direct Phonetic Segmentation

As the TIMIT database contains transcriptions at the phone level, it enabled us to evaluate firstly the accuracy of phonetic segmentation with maximally precise inputs of HMM-based forced alignment. In fact, it means the optimum input of forced-alignment with 100% correct phonetic content when no phone needs to be recognized and PER is equal to 0 %. Obtained results are in the Table 3. Similarly, as in several other works (e.g. [7] or [16]), the best results were obtained for the simplest monophone AM, for both the core and complete test sets. Slightly lower accuracy of triphone- and DNN-based AMs might be caused due to the fact that input features are taken from larger context, which yields to higher uncertainty in determination of a boundary position. Furthermore, speaker dependent AMs are probably estimated with smaller accuracy due to the limited amount of data per speaker in TIMIT corpus.

Concerning the monophone AM, we looked for its optimized setup. Same as in other published works' [7], it was confirmed that smaller amount of Gaussian mixtures per state gave better results. The best ones were achieved for 2 mixtures per state, see Table 4. The number in acronyms mono144, mono288, etc. in Tables 3 and 4 represents the number of Gaussian components in whole HMM, e.g. 288 means 288 components for 2 mixtures per state, 3 emiting states per each monophone, and 48 phones in given HMM ($2 \times 3 \times 48$).

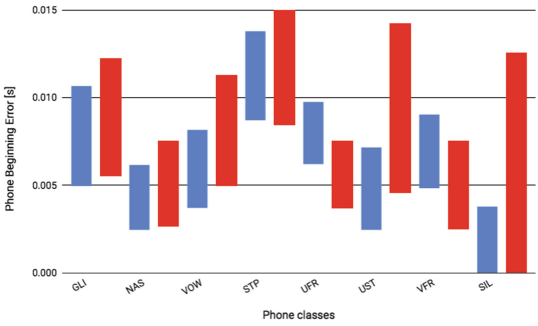

Fig. 1. Phone Beginning Error *PBE* for particular phone classes: blue - monophone system, red - DNN-based system. (Color figure online)

Table 3. Results of direct phonetic segmentation, $PER = 0$, $PCorr = 100$.

	CORE SET				COMPLETE SET			
	5 ms	10 ms	20 ms	30 ms	5 ms	10 ms	20 ms	30 ms
mono	**29.16**	**52.79**	**83.08**	**93.00**	**29.00**	**52.71**	**82.79**	**92.63**
tri1	27.80	51.21	81.69	92.82	27.84	50.89	81.40	92.12
tri2	27.40	49.55	79.72	91.45	27.10	48.96	79.27	90.91
tri3	27.42	49.34	79.18	91.24	27.18	48.74	78.41	90.36
dnn	27.73	48.87	78.84	90.77	27.11	48.49	78.32	90.09

Table 4. Optimization of monophone AM for direct phonetic segmentation ($PER = 0$, $PCorr = 100$).

	CORE SET				COMPLETE SET			
	5 ms	10 ms	20 ms	30 ms	5 ms	10 ms	20 ms	30 ms
mono144	31.05	54.57	82.51	92.17	31.37	54.67	81.90	91.73
mono288	**31.68**	**55.80**	84.70	**93.79**	**32.02**	**56.39**	**84.55**	**93.11**
mono432	30.45	54.73	**84.74**	93.74	31.03	55.32	84.46	93.06
mono720	29.76	53.50	83.53	93.35	29.95	53.70	83.48	92.99
mono1008	29.16	52.79	83.08	93.00	29.00	52.71	82.79	92.63
mono1440	28.18	51.50	81.80	92.82	28.13	51.31	81.80	92.30

Finally, the distribution of values of *PBE* for particular phone classes is presented in Fig. 1. Particular bars describe distribution of *PBE* determined by percentiles 0.25 and 0.75 and significantly worse results are observed for DNN system, however, significant increase of an error can be observed mainly for silence while deterioration within phone classes is not so critical.

3.3.2 Phonetic Segmentation with Pronunciation Variability

The second analysis describes the phonetic segmentation when exact phone sequence is not available and phonetic content is obtained from a pronunciation lexicon. It is the most frequent approach for obtaining phonetic content of an utterance, however, the core issue is how well the variability of pronunciation is covered in the lexicon and how the proper choice of word pronunciation variant influences the accuracy of phonetic segmentation.

Table 5. Phonetic segmentation with canonic lexicon.

		PER	*PCorr*	5 ms	10 ms	20 ms	30 ms
CORE	mono	32.58	71.43	23.94	43.54	72.39	85.82
	dnn	31.88	71.45	23.67	40.14	65.28	80.60
	mono288-dnn	31.88	71.45	25.78	45.37	72.01	84.54
COMPLETE	mono	31.15	72.28	23.92	43.23	72.34	85.78
	dnn	30.52	72.28	23.43	40.32	65.83	80.59
	mono288-dnn	30.52	72.28	26.39	46.38	72.71	84.93

Table 6. Phonetic segmentation with TIMIT-variant lexicon.

		PER	*PCorr*	5 ms	10 ms	20 ms	30 ms
CORE	mono	12.24	89.69	28.77	51.61	82.26	92.58
	dnn	**9.58**	92.03	27.64	48.55	78.03	90.05
	mono288-dnn	**9.58**	92.03	**31.28**	**54.94**	**83.81**	**93.09**
COMPLETE	mono	12.06	89.62	28.82	52.11	82.25	92.30
	dnn	**10.00**	92.06	27.16	48.28	77.73	89.55
	mono288-dnn	**10.00**	92.06	**31.91**	**55.98**	**84.17**	**92.93**

We realized the experiments with 3 pronunciation lexica: the first lexicon contained just *canonic pronunciations*, the second one contained all *pronunciation variants* realized by speakers in TIMIT corpus, and the third one was based on merging previous two lexica. Obtained results are shown in Tables 5, 6 and 7 and significant decrease of *PER* was observed when lexicon contained pronunciation variants. Further, the usage of more advanced AM (DNN-based one) contributed to further decrease of achieved *PER* below 10%. Consequently, it means the increase of *PCorr*, i.e. more than 92% of all phones were correctly identified, however, the accuracy of boundary determination slightly decreased when DNN-based system was used. On the other hand, when the recognized phone sequence is realigned with optimized monophone system with 288 Gaussian components (acronym mono288-dnn), both the best *PER* and boundary placement accuracy were achieved [15].

Table 7. Phonetic segmentation with canonic lexicon extended by TIMIT variants.

		PER	*PCorr*	5 ms	10 ms	20 ms	30 ms
CORE	mono	12.43	89.48	28.79	51.69	82.25	92.64
	dnn	**9.76**	91.88	27.65	48.51	77.99	90.00
	mono288-dnn	**9.76**	91.88	**31.33**	**55.00**	**83.84**	**93.12**
COMPLETE	mono	12.40	89.28	28.83	52.08	82.25	92.31
	dnn	**9.28**	92.17	27.12	48.22	77.63	89.44
	mono288-dnn	**9.28**	92.17	**31.92**	**55.97**	**84.14**	**92.93**

3.3.3 Pronunciation Recognition

In the end, we analyzed the correctness of pronunciation variant selection mentioned above. In fact, it was already quantified a little by the decrease of *PER* described in previous section, however, for many words we had a rather high amount of pronunciation variants and the ability of the selection of correct pronunciation variant could be very important feature of such a system. From the results described in Table 8, we can observe significant decrease of *PronER* (Pronunciation Error rate) when more advanced acoustic modeling and the lexicon covering pronunciation variants are used. The best results were obtained with DNN-based system, we observed significant decrease of *PronER*; 76.34% were obtained for basic monophone system and CORE test set, while 31.89% were achieved for DNN-based system. The contribution of GMM-HMM systems with triphone-based models was proven too. The same trend in obtained results was observed also for COMPLETE set.

Table 8. Pronunciation variant recognition.

		canonic		timit		canonic+variants	
		PER	PronER	PER	PronER	PER	PronER
CORE	mono	32.58	76.34	12.24	39.48	12.43	40.18
	tri1	32.80	76.28	11.49	37.82	11.74	38.46
	tri2	32.55	76.28	11.16	35.97	11.31	36.54
	tri3	32.46	76.28	10.24	33.48	10.42	34.06
	dnn	31.88	**76.34**	9.58	**31.44**	9.76	**31.89**
COMPLETE	mono	31.15	74.22	12.06	40.39	12.40	41.44
	tri1	31.79	74.21	11.89	37.06	11.45	37.87
	tri2	31.45	74.22	11.17	35.77	11.00	36.60
	tri3	31.30	74.21	10.75	33.82	10.23	34.56
	dnn	30.52	**74.22**	10.00	**31.46**	9.28	**32.19**

4 Conclusions

The implementation of HMM-based phonetic segmentation realized by KALDI toolkit was described in this paper commonly with the analysis of an contribution of various acoustic modeling to final accuracy of phone-boundaries determination. The evaluations were performed with TIMIT database and they proved the contribution of advanced acoustic modeling for the choice of proper pronunciation variant. We achieved more than 92% correctness of phone recognition within forced-alignment with DNN-HMM system together with the improvement of phone boundary placement realized in the second step by optimized monophone GMM-based systems; 83.84% of phone beginning boundaries were determined with the error smaller than 20 ms, for the error smaller than 30 ms it was 93.12%. These results were obtained without any further boundary correction, as it is not currently required by our application as well as it is related to results obtained without any boundary refinement and published by other authors. For the purpose of pronunciation variability modeling, the lexicon covering pronunciation variants of particular words among TIMIT speakers was created on the basis of phonetic transcriptions available in this corpus.

Acknowledgments. The research described in this paper was supported by internal CTU grant SGS17/183/OHK3/3T/13 “Special Applications of Signal Processing”.

References

1. CMUSphinx: Open source speech recognition toolkit. http://cmusphinx.github.io
2. Brunet, R.G., Murthy, H.A.: Pronunciation variation across different dialects for English: a syllable-centric approach. In: 2012 National Conference on Communications (NCC) (2012)
3. Garofolo, J.S., et al.: TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web download. Linguistic Data Consortium, Philadelphia (1993)
4. Ghoshal, A., Povey, D.: Sequence-discriminative training of deep neural networks. In: Proceedings of the INTERSPEECH, Lyon, France (2013)
5. Kahn, A., Steiner, I.: Qualitative evaluation and error analysis of phonetic segmentation. In: 28. Konferenz Elektronische Sprachsignalverarbeitung, Saarbrücken, Germany, pp. 138–144 (2017)
6. Lee, K.F., Hon, H.W.: Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Audio Speech Lang. Process. **37**(11), 1641–1648 (1989)
7. Matoušek, J., Klíma, M.: Automatic phonetic segmentation using the KALDI toolkit. In: Ekštein, K., Matoušek, V. (eds.) TSD 2017. LNCS (LNAI), vol. 10415, pp. 138–146. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_16
8. Matoušek, J., Tihelka, D., Psutka, J.: Experiments with automatic segmentation for Czech speech synthesis. In: Matoušek, V., Mautner, P. (eds.) TSD 2003. LNCS (LNAI), vol. 2807, pp. 287–294. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39398-6_41

9. Mizera, P., Pollak, P., Kolman, A., Ernestus, M.: Impact of irregular pronunciation on phonetic segmentation of Nijmegen corpus of casual Czech. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2014. LNCS (LNAI), vol. 8655, pp. 499–506. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10816-2_60
10. Nouza, J., Silovský, J.: Adapting lexical and language models for transcription of highly spontaneous spoken Czech. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 377–384. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15760-8_48
11. Peddinti, V., Wang, Y., Povey, D., Khudanpur, S.: Low latency acoustic modeling using temporal convolution and LSTMs. IEEE Signal Process. Lett. **25**(3), 373–377 (2018)
12. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, ASRU 2011 (2011)
13. Rendel, A., Sorin, A., Hoory, R., Breen, A.: Toward automatic phonetic segmentation for TTS. In: Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, pp. 4533–4536 (2012)
14. Rybach, D., et al.: The RWTH Aachen university open source speech recognition system. In: Proceedings of Interspeech 2009 (2009)
15. Stolcke, A., Ryant, N., Mitra, V., Yuan, J., Wang, W., Liberman, M.: Highly accurate phonetic segmentation using boundary correction models and system fusion. In: Proceedings of ICASSP, Florence, Italy (2014)
16. Toledano, D.T., Gómez, L.A.H., Grande, L.V.: Automatic phoneme segmentation. IEEE Trans. Speech Audio Process. **11**(6), 617–625 (2003)
17. Young, S., et al.: The HTK Book, Version 3.4.1. Cambridge (2009)
18. Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., Wang, W.: Automatic phonetic segmentation using boundary models. In: Proceedings of INTERSPEECH, Lyon, France, pp. 2306–2310 (2013)