A Phonetic Segmentation Procedure Based on Hidden Markov Models

Edvin Pakoci[1], Branislav Popović[1(✉)], Nikša Jakovljević[1], Darko Pekar[2], and Fathy Yassa[3]

[1] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
{edvin.pakoci,bpopovic,jakovnik}@uns.ac.rs
[2] AlfaNum Speech Technologies, Novi Sad, Serbia
darko.pekar@alfanum.co.rs
[3] Speech Morphing Inc., Campbell, CA, USA
fathy@speechmorphing.com

Abstract. In this paper, a novel variant of an automatic phonetic segmentation procedure is presented, especially useful if data is scarce. The procedure uses the Kaldi speech recognition toolkit as its basis, and combines and modifies several existing methods and Kaldi recipes. Both the specifics of model training and test data alignment are explained in detail. Effectiveness of artificial extension of the starting amount of manually labeled material during training is examined as well. Experimental results show the admirable overall correctness of the proposed procedure in the given test environment. Several variants of the procedure are compared, and the usage of speaker-adapted context-dependent triphone models trained without the expanded manually checked data is proven to produce the best results. A few ways to improve the procedure even more, as well as future work, are also discussed.

Keywords: Kaldi · Phonetic segmentation · Hidden Markov models

1 Introduction

In recent years, there is an evident increase in the amount of available multimedia data including speech. This data is interesting for research in social sciences, as well as for speech technologies. These studies usually require audio content and the corresponding phonetic transcription synchronized with it. Manual alignment of audio and text data is very laborious and expensive (30 s of audio data requires about an hour of manual work [1]), thus many automatic and semi-automatic procedures have been developed.

All these procedures can be classified into two broad groups depending on whether or not additional acoustic information about phone identities are used. The first group is comprised of methods which for phonetic segmentation use the information contained in the given audio signal, and order of phones in the corresponding phoneme sequence. These methods are referred to as text-independent or linguistically unconstrained segmentation methods. They exploit

A. Ronzhin et al. (Eds.): SPECOM 2016, LNAI 9811, pp. 67–74, 2016.
DOI: 10.1007/978-3-319-43958-7_7

the fact that sudden changes in speech signal characteristics usually coincide with phone boundaries. Exceptions of this rule are plosives/affricates consisted of occlusion and explosion/friction parts as well as transitions between successive vowels or vowels and semi-vowels [2]. These changes are usually detected in the spectral or cepstral domain [3–5]. Additionally, the level of feature similarity can be exploited in segmentation as in [6]. An advantage of these methods is their independence from language, but the accuracies of obtained phone boundaries are significantly poorer compared to accuracies of text-dependent methods.

Text-dependent or linguistically constrained segmentation methods align audio signal with corresponding phone sequence using phone models similar to those used in the automatic speech recognition (ASR) task. The dominant approach to phone modeling is hidden Markov models (HMMs), and interesting results are obtained in [7–13], among others. Besides HMMs, dynamic time warping [14] and artificial neural networks - ANNs [15,16] are used as well.

The width of analysis frames varies from 10 ms up to 30 ms, and frame shift varies from 1/5 of the frame width up to whole frame width. There are some variations of the extracted features in existing methods, but most of them include 12–14 mel-frequency cepstral coefficients (MFCCs), normalized energy and their first and second order time derivatives. In some studies, the set of features additionally includes a spectral variation function [17], perceptual loudness, measure of periodicity [9] and fundamental frequency (f_0) contour [15]. The basic modeling units can be monophones, triphones or tied-state triphones. The choice between those depends primarily on the size of the training corpus. Sometimes the improvement in alignment accuracy, which is usually obtained with context-dependent modeling units, is not sufficient to justify the increase of duration of the training procedure. Since the objective function for estimation of HMM parameters does not involve accurate position of phone boundaries after alignment, additional boundary refinement is possible. It is usually based on principles exploited in linguistically unconstrained methods [10,17] or using trained GMM or ANN models for boundaries [15,16]. The proposed procedure belongs to the group of linguistically constrained methods based on HMMs in case of scarce data. The procedure is tested on several databases in English whose description is given in Sect. 2. Detailed description of the procedure is presented in Sect. 3, and results of evaluation with discussion in Sect. 4. Section 5 concludes the paper.

2 Speech Corpora

Appen "USE_ASR001" database [18] of natural English (US) speech in studio quality, resampled to 16 kHz, 16 bits per sample, mono PCM, was used in our research. The database contains more than 80000 utterances (almost 7 GB of data, 100 male and 101 female speakers), or approximately 41 h of speech and 20 h of silence segments, and it is transcribed in SAMPA format. Only a small part of the database (in further text referred to as the bootstrapped part), containing around 35 min of speech and 15 min of silence segments, was manually labeled. This part of the database consists of short cropped audio files

(around 1900) containing only about 2 to 3 words, which are selected in a way to cover all phone pairs (one after the other) which exist in the source database among all the given speakers. The extraction of audio segments is done automatically using information from initial forced alignment using flat-start models. These phone boundaries are manually checked and corrected by trained annotators. It has been shown that by applying manual alignments on a part of the database throughout the procedure, significant improvement can be obtained in comparison to flat-start training (e.g. [9]). Additionally, in our procedure the bootstrapped part of the database was artificially expanded several times, by modifying pitch and duration (i.e., by applying spectral warping and tempo modifications). It was feared that such a procedure could not provide sufficient data variability, so it was only applied in a limited amount. The original tempo was increased or decreased by 10 % and 20 % and the original pitch was increased (for males) or decreased (for females) by 1, 2 or 3 semitones. Male speakers whose pitch was increased by 1 and 2 semitones, along with the original unmodified male speakers, and female speakers whose pitch was decreased by 3 semitones, were used for the training of specialized male models. The bootstrapped part of the database (along with all the mentioned extensions) was then additionally doubled by marking all the words and phonemes as damaged in the copied instance. This was done in order to provide minimal number of samples needed to train the damaged phoneme models - they were needed primarily since the starting and ending phones in cropped segments had to be marked as damaged, as these are not full sentences by themselves, but segments not necessarily surrounded by silence. Therefore, the bootstrapped part of the database was increased 40 times for male speakers ([male + male pitch $\{+1, +2\}$ + female pitch $\{-3\}$] $\times$ tempo $\{-20, -10, 0, +10, +20\} \times 2$ for damaged phonemes). A special characteristic of our training was that the phone boundaries on the whole bootstrapped set (expanded) were kept fixed during the entire procedure, so they could have a greater influence on the accuracy of alignment in the remainder of the database.

Our test database, on which the results presented in this paper were calculated, included an array of phonetically rich utterances, spoken by 3 male speakers - Sean, Doug and Ben - from completely independent single-speaker databases, provided by Speech Morphing Inc. All the utterances for testing were manually labeled, so that automatically aligned phone boundaries could be compared to them. Sean's test database contains 50 utterances (around 1800 phonemes), which added up to 2 min 13 s of speech and 33 s of silence. Ben's test database contains 43 slightly longer utterances, 4 min 17 s of speech and 1 min 5 s of silence in total (around 2700 phonemes). Finally, Doug's test database contains 50 utterances - 2 min 28 s of speech and 23 s of silence (around 1800 phonemes). At the time of tests, no similar female databases were at our disposal. Nevertheless, the obtained results confirm our previous assertions and they were highly comparable among all test databases, as shown by the experiments (see Sect. 4).

3 Segmentation Procedure

The complete training procedure has been done using the Kaldi speech recognition toolkit [19] and modified Kaldi recipes. Inputs were Kaldi data files created using an input lexicon (in SAMPA format), which included all needed words and their pronunciations with multiple alternative pronunciations in some cases, as well as utterance transcriptions with marked speaker identifiers. Model topologies were initialized to 3 states for non-silence phones, and 5 states for silence phones, with a possibility to skip one state at a time (for a minimum of 3 states). For decoding purposes, as in all Kaldi training procedures, a lexicon FST (i.e., finite state transducer) is created based on the input lexicon. In our procedure, this FST is modified by adding alternative arcs for all arcs that have a vowel as their input label, and it concerns vowel stress - if the vowel is stressed, alternative arc with the unstressed version of the same vowel as input label is created, and vice versa. Also, for the arc containing the optional silence between and after each spoken word, an arc with optional glottal stop is created as an alternative. This was done because a lot of places in the database were identified where the gap between words includes rather a glottal stop then something than can be considered as a silence (which would lead to a "dirty" silence model).

The feature vectors include energy and 14 MFCCs, calculated by using a filter bank of 26 overlapping triangular windows, along with their first and second order time derivatives. They were extracted on 30 ms frames, with 7 ms frame shift. Multiplication coefficient of 0.33 was additionally applied to static MFCCs (excluding energy) to bring their value variability closer to that of energy. On the other hand, delta and delta-delta energy values were multiplied by coefficient 20, to effectively change their dynamic range. The mentioned coefficient values were concluded to be appropriate through several previous tests and extracted feature values analysis. No cepstral mean or variance normalization is performed, as the training type (explained below) makes it unnecessary.

The first stage of model training, which is the training of monophones on bootstrapped data set only, comes next. This step included manual alignments as the starting point and a 10 iterations of model and alignment reestimation. After each internal alignment, phone boundaries were reset to manually-given positions, but the within-phone frames per state distributions were saved. Output was the final monophone model set with 1000 Gaussians in total. Afterwards, these models were used to create the initial context-dependency tree for the speaker-adapted training (SAT) step, and to produce initial alignments for the rest of the database. The first pass of SAT started from the aforementioned alignments, i.e., there is no equidistant initial alignment at all. The bootstrapped data set is still used, alongside the whole regular database. The context-dependency tree which is created here has a goal of 1500 leaves (i.e., states). Next, initial fMLLR transforms are calculated using initial alignments, producing a diagonal transform matrix for each of the speakers in the database. Then 10 iterations of model reestimation follow, with periodic internal alignments and fMLLR transform matrices updates, ending with a goal of 4500 Gaussians for final models. Manually set boundaries are also forced throughout the stage

(in the bootstrapped part of the database). In the end, the so-called "alignment model" is created - it is computed with speaker-independent (SI) features, but matches Gaussian-for-Gaussian with the speaker-adapted model.

The training procedure ends with the final SAT training pass. The alignment model is first used to align the whole database. Also, a new tree is created, slightly more complex with 2500 leaves, using these new, better alignments. The rest of the stage is very similar to the previous stage, with 12 iterations and a goal of 7000 Gaussians. It outputs the final SAT model set, as well as the final alignment model set. In all internal alignments a large decoding beam is used, to prevent potentially important tokens from being discarded in the early stages of utterance decoding. All the selected numbers - of iterations, states and Gaussians - have shown the best performance on a validation set (a part of the bootstrapped set) during exhaustive testing where these numbers were varied. Now that the models are ready, they can be used to align the given test data. The start is the same as for the training – there are given transcriptions matched to an audio file name each with marked speaker identifiers, and a lexicon containing all the words in transcriptions with possible pronunciations. These are converted to appropriate Kaldi data files. The procedure setup has to be the same as for model training - number of states for certain phones, list of used phones, MFCC and energy extraction specifications. Firstly, lexicon FST is created and modified the same way as in training, which is followed by static feature extraction. Delta and delta-delta features are added later on the fly. Decoding graphs are created from lexicon FST, provided models and corresponding tree. This is followed by first-pass alignment using SI features and the alignment model, the output of which is used to estimate fMLLR transforms, producing a diagonal matrix for each speaker in the test database, used to transform features (on the fly). In the end, the final alignment is performed, using transformed features and provided SAT models. The results are phone alignments within a label file.

4 Results and Discussion

Our test data sets include exclusively male speakers, so the results were obtained using male models. Several experiments were conducted. First, the possibility of using simple monophone models to align test data directly using just SI features is examined [9]. This of course shortens the training procedure a lot, but monophone models may not be precise enough. Then, the described procedure with triphones and SAT is evaluated. Both of these experiments are performed both by using the basic bootstrapped set (without tempo and pitch modifications), and the fully extended bootstrapped set. All the given results were obtained by calculating the difference between the phone start times in the manually labeled set and in the automatically obtained labels, then putting each of those numbers in the appropriate category based on the difference, e.g. up to 10 ms, up to 20 ms, and so on. In the special cases of inserted and deleted optional silences and glottal stops, they are instead compared to the previous phone in the manually marked database (if inserted) or automatic labels (if deleted).

Results for monophones with the extended bootstrapped set are given in Fig. 1 (left). Percentages for phones within 10 ms on all test sets are around 60 %, with more than 80 % within 20 ms, around 90 % within 30 ms and 95 % within 50 ms, and around 5 % of outliers. Outliers mostly include silences or phones after silences, especially if the neighboring phone is a plosive, affricate or a silent fricative. If silences and their adjacent phones are excluded from the results, around 85 % of phone boundaries fall within 20 ms of manual ones. The remaining outliers are mostly boundaries between two plosives, two similar vowels and finally borders between some vowels and lateral 'L' or approximant 'R', which is not that surprising as these borders are hard to put in the correct place even by hand (at most times there is actually no clear border). These kinds of outliers appear in other experiments as well. As for context-dependent triphones trained with the SAT procedure, the results are given in Fig. 1 (right). For the percentage of phones within 20 ms of manual boundaries, a 4–7% improvement was obtained at average, which is a lot when talking about segmentation quality. After excluding silence borders, these improve to over 90 %. The usage of triphones and SAT is justified, even considering the longer training.

The results with the non-extended bootstrapped set are shown in Fig. 2. The results are better then with artificial extension. It can be assumed that the artificial extension of the bootstrapped set results in significant feature dispersion which could not be covered with the monophone models using the same target

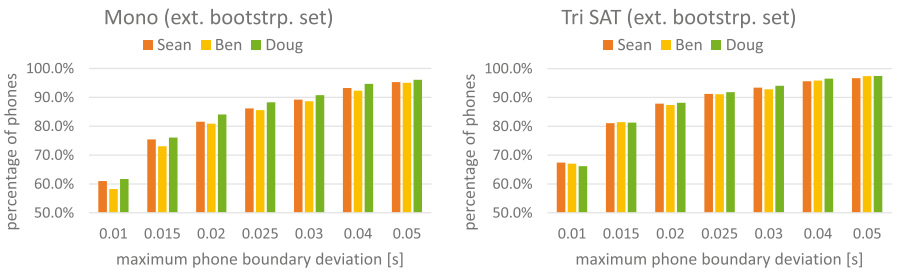

Fig. 1. Results for extended bootstrapped set.

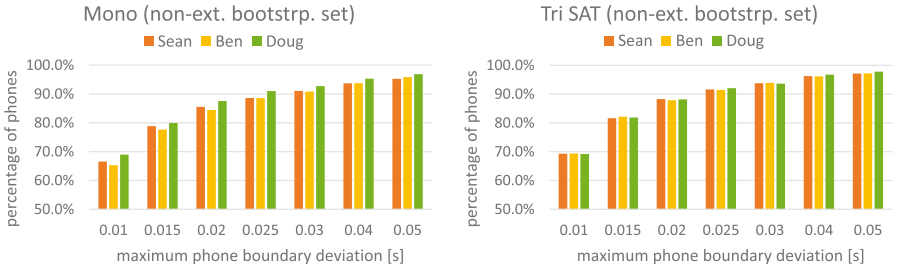

Fig. 2. Results for non-extended bootstrapped set.

number of Gaussians. On the other hand, in case of tied-state triphones it is largely compensated by fMLLR. It will be a subject of further research.

5 Conclusion and Further Directions

In this paper, a novel approach to automatic phone segmentation of an arbitrary speech database in case of scarce data is presented. It is concluded that context-dependent SAT models produce best and most stable overall results, but monophones are not too far behind, if procedure speed is of more concern. Artificial extension of the manually labeled part of the database is examined as well, and it was not proven to improve the results, at least if used in the way described. In the near future, more experiments will be done with versions of the expansion procedure which will conclude what exactly went wrong here. For now, it is assumed that either bad modification parameters are chosen, or the expansion went too far (the part with manual boundaries became too significant compared to the rest of the database). Future work will also include training parameter variations, other speech databases (including other languages as well), and finally improving the alignment analysis tool to get even more data which can help with pointing in the right direction. The greatest value of the described procedure is that the obtained correctly aligned speech databases can be used relatively quickly and successfully for any given application.

Acknowledgments. This research was supported in part by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under Grant No. TR32035. The authors are grateful to the company "Speech Morphing, Inc." from Campbell, CA, USA, for providing the speech corpora for the experiments.

References

1. Brognaux, S., Roekhaut, S., Drugman, T., Beaufort, R.: Train&Align: a new online tool for automatic phonetic alignment. In: Spoken Language Technology Workshop (SLT), pp. 416–421. IEEE Signal Processing Society (2012)
2. Scharenborg, O., Ernestus, M., Wan, V.: Segmentation of speech: child's play? In: 8th Annual Conference of the International Speech Communication Association (INTERSPEECH), Antwerp, pp. 1953–1956 (2007)
3. Esposito, A., Aversano, G.: Text independent methods for speech segmentation. In: Chollet, G., Esposito, A., Faundez-Zanuy, M., Marinaro, M. (eds.) Nonlinear Speech Modeling. LNCS (LNAI), vol. 3445, pp. 261–290. Springer, Heidelberg (2005)
4. Leow, S.J., Chng, E.S., Lee, C.H.: Language-resource independent speech segmentation using cues from a spectrogram image. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, pp. 5813–5817 (2015)
5. Priyadarsini, S., Kumar, A.: Automatic speech segmentation in syllable centric speech recognition system. J. Speech Technol. **19**(1), 9–18 (2016)

6. Almpanidis, G., Kotti, M., Kotropoulos, C.: Robust detection of phone boundaries using model selection criteria with few observations. IEEE Trans. Audio Speech Lang. Process. **17**(2), 287–298 (2009). IEEE Signal Processing Society
7. Bigi, B.: SPPAS: a tool for the phonetic segmentations of speech. In: 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, pp. 1748–1755 (2012)
8. Boeffard, O., Charonnat, L., Le Maguer, S., Lolive, D., Vidal, G.: Towards fully automatic annotation of audio books for TTS. In: 8th International Conference on Language Resources and Evaluation (LREC), Instanbul, pp. 975–980 (2012)
9. Brognaux, S., Drugman, T.: HMM-based speech segmentation: improvements of fully automatic approaches. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(1), 5–15 (2016). IEEE Signal Processing Society
10. Hoffmann, S., Pfister, B.: Fully automatic segmentation for prosodic speech corpora. In: 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), Makuhari, pp. 1389–1392 (2010)
11. Hoffmann, S., Pfister, B.: Text-to-speech alignment of long recordings using universal phone models. In: 14th Annual Conference of the International Speech Communication Association (INTERSPEECH), Lyon, pp. 1520–1524 (2013)
12. Matoušek, J.: Automatic pitch-synchronous phonetic segmentation with context-independent HMMs. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 178–185. Springer, Heidelberg (2009)
13. Stan, A., Mamiya, Y., Yamagishi, J., Bell, P., Watts, O., Clark, R.A.J., King, S.: ALISA: an automatic lightly supervised speech segmentation and alignment tool. J. Comput. Speech Lang. **35**, 116–133 (2016)
14. Adell, J., Bonafonte, A., Gomez, J., Castro, M.: Comparative study of automatic phone segmentation methods for TTS. In: 30th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, pp. 309–312 (2005)
15. Toledano, D., Gomez, L., Grande, L.: Automatic phonetic segmentation. IEEE Trans. Speech Audio Process. **11**(6), 617–625 (2003). IEEE Signal Processing Society
16. Wang, L., Zhao, Y., Chu, M., Zhou, J., Cao, Z.: Refining segmental boundaries for TTS database using fine contextual-dependent boundary models. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Montreal, pp. 641–644 (2004)
17. Brugnara, F., Falavigna, D., Omologo, M.: Automatic segmentation and labeling of speech based on hidden Markov models. J. Speech Commun. **12**(4), 357–370 (1993)
18. Appen, Product Catalog. http://catalog.appenbutlerhill.com/
19. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Veselý, K.: The kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 1–4. IEEE Signal Processing Society (2011)