Fully Automatic Segmentation for Prosodic Speech Corpora

Sarah Hoffmann, Beat Pfister

Speech Processing Group, ETH Zürich, Switzerland
{hoffmann,pfister}@tik.ee.ethz.ch

Abstract

While automatic methods for phonetic segmentation of speech can help with rapid annotation of corpora, most methods rely either on manually segmented data to initially train the process or manual post-processing. This is very time-consuming and slows down porting of speech systems to new languages. In the context of prosody corpora for text-to-speech (TTS) systems, we investigated methods for fully automatic phoneme segmentation using only the corpora to be segmented and an automatically generated transcription. We present a new method that improves the performance of HMM-based segmentation by correcting the boundaries between the training stages of the phoneme models with high precision. We show that, while initially aimed at single speaker corpora, it performs equally well for multi-speaker corpora.

Index Terms: speech synthesis, segmentation, boundary correction

1. Introduction

Phonetically labelled speech corpora are widely used in speech processing and, consequently, automatic phonetic segmentation has been widely researched. However, improvement of segmentation methods has focused on reducing the amount of required manually segmented data, less on completely eliminating human intervention. In the context of our multi-lingual text-to-speech system SVOX, we require annotated monolingual corpora of each supported language for the training of the duration models of the prosodic component [1]. To be able to use the system also with less widely used languages, the segmentation process must not rely on the availability of any manually segmented data for the language. Instead, it should use only information extracted from the corpora to be segmented. Prosody corpora are normally recorded by professional speakers, so they contain very few slips and the phones are spoken very consistently. A transcription of the corpus can be obtained in an automatic way from the orthographic text presented to the speaker with the help of the syntactic component of the TTS system. Such an automatic transcription does not necessarily reflect the exact content of the speech signal as the speaker may use different pronunciation variants or rapid speech phenomena like assimilation may occur.

In summary, we require a segmentation process that is able to segment a corpus using only corpus data and a transcription. It must not use any manually annotated data, neither for training of statistical segmentation models nor during the segmentation process itself. In addition, it must be robust even in the presence of transcription errors.

Existing segmentation methods can be grouped into those that rely on statistical models and those that are text-independent, i.e. those that try to estimate phone boundaries from information in the speech signal only. Examples for the latter are [2], who proposes boundary detection based on jumps in the feature distance of acoustic features, and [3], who estimates boundaries by analysing energy changes in different spectral bands. While these methods report a very high recall rate, they suffer from the problem of over-segmentation. The problem remains to correctly assign the transcription.

Statistical methods can be used if the models can be trained directly on the corpus to be segmented. Most widely used are HMMs in forced-alignment mode. The phone models can be flat-start initialized and then trained directly on the corpus. However, the resulting segmentation is less accurate than when manually segmented training material is used (see e.g. [4]). [5] reported better results by using a hierarchical method, segmenting into broader phonetic classes first.

Various methods have been proposed to improve HMM-based segmentation. The largest group uses statistical models, for example, [6] proposes regression trees, [7] applies statistical error models followed by a boundary correction using neural networks. These approaches improve boundaries well but are not suitable for a fully automated segmentation approach, as the training of these statistical correction models again requires manually segmented training material. [8] proposes a text-independent method for boundary correction. The spectral discontinuity in the signal is computed and the boundaries are moved to the closest local maximum. They report a very high error rate for the correction and propose to impose empirically determined limits on how much a boundary can be moved by the correction process, again requiring human intervention.

We propose a boundary-correction algorithm similar to [8]. It relies on the local feature distance between adjacent phones and allows to predict the correct boundary with a much smaller error rate. In addition, we propose to embed the correction into the HMM training process to improve the quality of the phone models. The following section gives an overview of the segmentation process, after which the processing of the automatic transcription and the boundary correction is discussed in detail. Finally, the method is evaluated on a prosody corpus and the TIMIT speech database.

2. Overview of the segmentation process

An important reason for the worse performance of flat-started phone models in comparison to models trained on manual segmentation is that an unrestricted iterative embedded training is performed on data that is badly balanced in terms of phonetic context. The frequency of phoneme pairs in the corpus is determined by the language of the corpus. If a phoneme occurs very frequently in the same context then this fact is 'learned' by the HMM by including part of the context in the model. During segmentation it results in a systematic error in the boundary placement. Table 1 shows the percentage of correct boundaries within a 5ms and 20ms deviation for a HMM-based segmen-

HMM training method	Max. deviation 5ms	20ms	Misaligned
isolated phoneme training	51.13	89.50	0.59
embedded training	44.67	87.87	0.56
flat-start init. + embedded tr.	41.96	85.36	0.46

Table 1: Segmentation performance for various HMM training methods trained on the TIMIT corpus. Percentage of correctly placed boundaries for different maximum deviations and of misaligned labels, i.e. labels not overlapping with those in the reference segmentation.

tation that uses manually segmented data and the performance after the same models have been further trained for 40 iterations using embedded training. The boundary correctness has decreased significantly after the embedded training.

To avoid this issue, we divide the segmentation process in two stages as depicted in Figure 1. The first stage uses the automatic transcription as a basis. The phone models are flat-start initialized and iteratively trained using embedded training. A segmentation is produced via forced alignment and, at the same time, optional silences, plosive pauses and glottal stops inserted. The boundaries of the resulting segmentation are corrected, resulting in a preliminary segmentation that is the base for the next stage.

In the second stage, the phone models are estimated from the preliminary segmentation using isolated-unit training only, thus the models retain the boundary information. Another segmentation is computed, the insertion of silences, glottal stops and plosive pauses is reevaluated and the boundaries are corrected giving the final segmentation. The second stage can be repeated, if required.

3. HMM-based segmentation

The objective of the HMM-based segmentation is to obtain a segmentation where the phones are placed as close as possible to their true position. Experiments showed that the most stable results are obtained with 5-state left-to-right linear models with features computed every 4ms over a 20ms window. This results in a minimum length of 20ms for each phone. The features are comprised of the commonly used set of the first 12 MFCC coefficients, log energy and their first-order derivatives. Only one Gaussian mixture was used as the targeted corpora consists of only one speaker speaking with high consistency. The HMM-based segmentation process was implemented with the HTK toolkit [9].

3.1. Extending the automatic transcription

The automatic transcription determines the phoneme set present in the final segmentation and thus the set of phone models. Each phoneme is mapped to one phone model with the exception of diphthongs and affricates. Diphthongs contain two stable regions, splitting them into two parts improves the boundary correction algorithm below. Affricates are handled like fricatives with plosive pauses in front.

Silences, glottal stops and plosive pauses require special handling. As the speaker has a high degree of freedom in how to realise them, they do not appear in the automatic transcription. They must be added to the segmentation for two reasons: first, they are required for the training of the prosodic models which should learn how they are realised. Second, they are very different from the rest of the phoneme set. Having separate models improves therefore both, HMM-based segmentation and the boundary correction.

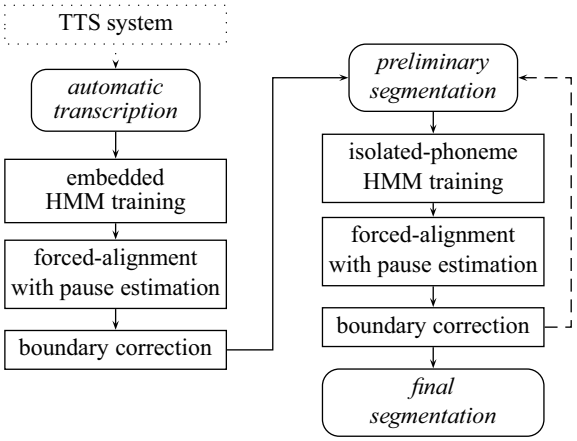

Figure 1: Complete segmentation process. The first stage on the left prepares a segmentation, which in the second stage on the right is used for training of HMMs.

The realisations of silences and glottal stops can be learned and determined during the HMM-based segmentation. Silence models are initially trained from the silences at the beginning and end of the utterances, glottal stop models from mandatory glottal stops as transcribed. If no mandatory glottal stops exist, a heuristic needs to be used. Experiments showed that the assumption that there is a glottal stop between each silence and vowel is sufficient. The actual placement of the phonemes is then determined by introducing variants during the forced-alignment process. Optional silences are added after each word and optional glottal stops in front of each syllable that begins with a vowel.

3.2. Plosive splitting

The plosive-pause model cannot be trained like the silence model. If plosive pauses are added in front of each plosive for the initial training, the left context of plosives is restricted to plosive pauses only. As explained in section 2, this causes parts of the pause to be trained into the plosive models. Therefore, the plosive models are trained to contain the plosive pause in the first stage and phones are subsequently split into pause and plosion.

The splitting algorithm has to take into account that plosives may not be realized because of fusion or elision. In that case only a silence will be seen. It can also be observed, albeit less frequently, that the plosive pause is omitted, for example, after fricatives.

As the boundary between pause and plosion is normally visible over the entire spectrum, it is sufficient to search for jumps in the overall spectral energy. For each frame in the plosive, the difference in the spectral energy over a 30ms window on both sides of the frame is computed. If the resulting function contains a peak and its maximum value is above 0, the phoneme is split at the position of the maximum. Otherwise, a heuristic is used to determine the nature of the phone: if its mean energy is below the mean energy of the silences in the utterance, a plosive pause is assumed, otherwise a plosion. Finally, where the splitting process created any succeeding plosive pauses, they are fused together.

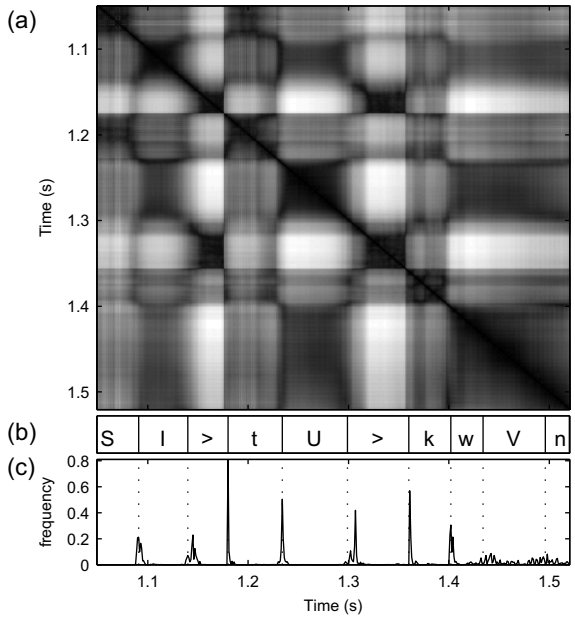

Figure 2: Feature-distance digram (a), reference segmentation (b) and distribution of boundary placement (c) for phrase 'she took one'.

4. Boundary correction

Phone boundaries can be considered as a transition phase between relatively stable centers of the phones and as such are very well visible in a frame-to-frame similarity matrix. Figure 2 (a) shows such a matrix for the phrase *she took one*. This property can be used to locally correct the boundary between two phones.

Given two adjacent phones p_i and p_{i+1}, the boundary is located between the perceptual centers of these phones. We call these centers *core frames*. Given the *core frames* c_i and c_{i+1}, the ideal boundary is placed such that all frames to the left are closer to the left core frame and all frames to right are closer to the right core frame. Such a boundary does not necessarily exist, so we compute the best boundary as follows: if $x_{c_i}, ..., x_{c_{i+1}}$ is the series of feature vectors between the two core frames, the confident boundary $\hat{b}_{i,i+1}$ from the left core frame and $\hat{b}_{i+1,i}$ from the right core frame are computed as

$$\hat{b}_{i,i+1} = b \quad | \quad D(x_{c_i}, x_b) \geq D(x_{c_{i+1}}, x_b) \wedge \forall\big[c_i < f < b : D(x_{c_i}, x_f) < D(x_{c_{i+1}}, x_f)\big] \tag{1}$$

$$\hat{b}_{i+1,i} = b \quad | \quad D(x_{c_i}, x_b) \leq D(x_{c_{i+1}}, x_b) \wedge \forall\big[b < f < c_{i+1} : D(x_{c_i}, x_f) > D(x_{c_{i+1}}, x_f)\big] \tag{2}$$

where D(x,y) is the Euclidean distance between two feature vectors x and y. Then, the final boundary between the phones p_i and p_{i+1} is

$$b_{i,i+1} = \left\lfloor \frac{\hat{b}_{i,i+1} + \hat{b}_{i+1,i}}{2} \right\rfloor \tag{3}$$

If the core frames are located in the true center of each phone, the boundaries will be placed as expected: on jumps in the distance function, if there is one, and equidistant from both core frames otherwise. However, the position of the core frames needs to be estimated from the placement of the phones in the imperfect preliminary segmentation. In order to evaluate how strongly the boundary function depends on the correct choice of the core frames, we computed the boundaries for any two frames from two adjacent phones from the reference segmentation. Figure 2(c) shows the result for the example phrase *she took one*. There is a clear preference of frames close to the reference boundaries. Therefore, choosing any frame that is close to the center of the phone is sufficient as core frame.

Assuming that the HMM-based segmentation has placed the phones in the right vicinity, we compute the frame that is most typical of the ones in the phone, i.e. that is closest to all other frames. The core frame c_i of a phone p_i with the preliminary boundaries b'_i and b'_{i+1} is then

$$c_i = \operatorname*{argmin}_{b'_i < f < b'_{i+1}} \Big(\operatorname*{median}_{b'_i < f' < b'_{i+1}} D(x_{f'}, x_f) \Big) \tag{4}$$

where $D(x, y)$ is again the Euclidean distance. A median is used instead of a mean to be more robust against noise in the signal.

In contrast to the HMM-based segmentation, where relatively coarse features are required to minimize segmentation errors, the boundary correction yields the best results with short-term features computed every 1ms over a 10ms window. Experiments show that perceptional features give slightly better results. The first 12 PLP coefficients as computed by the HTK toolkit together with normalized log energy are used.

5. Evaluation

The segmentation process was evaluated against a German prosody corpus recorded in our group and against the TIMIT Continuous Speech Corpus ([10]). We evaluated the precision of boundary placement on the one side, and the robustness of the segmentation on the other side, the latter by evaluating the number of misaligned labels, that is, labels that have no overlap in time with the corresponding labels from the reference segmentation.

5.1. Prosody corpus segmentation

The German prosody corpus consists of 186 sentences spoken by one male professional speaker sampled at 16kHz. A manual segmentation of the entire corpus was available from a previous project. The automatic transcription of the corpus is very close to the manual transcription. Differences are mostly related to optional glottal stops.

The corpus is very small for the training of the HMM. The phone models from the first stage generalize badly, which causes 0.9% of the phones to be misaligned. The boundary correction allows better models to be trained in the second stage.

	Max. deviation			Mis-
stage	5ms	10ms	20ms	aligned
HMM (1st stage)	40.23	66.61	86.61	0.93
HMM (2nd stage)	46.39	71.58	89.06	0.66
with bnd. correction	53.03	72.04	89.14	0.43
10th iteration	53.47	72.70	89.45	0.29

Table 2: Segmentation of German prosody corpus. Percentage of correctly set boundaries for different maximum deviations after HMM-based segmentation in 1st and 2nd stage, after correction in 2nd stage and after 10th iteration of 2nd stage.

That is why the iteration of the second stage does not suffer from diverging boundaries like embedded HMM training without correction does. The iteration process further reduces the number of misalignments by 47% after 10 iterations.

The boundary correctness of 89% within a 20ms deviation is close to the performance of similar HMMs trained with manually segmented data reported by e.g. [4] or [6]. Segmentation of a larger corpus is expected to perform equally well. With a correctness of 53% for a 5ms maximum deviation the boundary correction clearly outperforms any HMM-based segmentation in terms of precision.

5.2. TIMIT corpus segmentation

In order to evaluate the robustness of the segmentation with respect to imprecise transcription, we also segmented the TIMIT speech corpus of American English. The corpus differs from the targeted prosody corpora in that it was recorded from multiple speakers without professional training speaking multiple dialects. We used all 1344 sentences from the test set spoken by 168 different speakers of 6 different dialect regions. To compensate for the resulting greater phoneme variety the phone models have been trained with 4 Gaussian mixtures. Otherwise the HMM configuration was the same as described in section 3.

We did two sets of experiments, the first using manual, the second automatic transcription. We used the standard TIMIT phoneme set as defined in the lexicon. Where the reference segmentation distinguished additional phoneme variants, they were mapped to their closest counterpart from the standard set.

The manual transcription was taken directly from the reference segmentation including silences, glottal stops and plosive pauses. The plosive splitting algorithm was applied nonetheless. The automatic transcription was produced with the TIMIT lexicon and the word list provided for each sentence. If there was more than one transcription for a word one was chosen randomly.

Table 3 shows the results. Boundary correction improves the segmentation by 2.3% relative for manual and 3.0% relative for automatic transcription for a 20ms maximum deviation. [8] reported only 1.8% relative improvement, but on a segmentation with a higher baseline. The results are similar to those achieved on the prosody corpus. That shows that the correction works equally well for corpora with a higher variety in segmental quality.

For both experiments, the number of misaligned labels is reduced after boundary correction, making the method very robust against transcription errors. Additional errors are introduced by the HMM-based segmentation in the second stage, though, most probably, because the more precisely trained models are less flexible in skipping over transcription deviations. Increasing the Gaussian mixtures to allow more variation counters the effect in the case of manual transcription but not for the automatic one. Introduction of variants may be necessary instead.

segmentation method	Max. deviation 5ms	10ms	20ms	Misaligned
Manual transcription				
HMM (1st stage)	41.96	67.57	85.36	0.46
HMM (2nd stage)	49.43	73.44	88.22	0.48
with bnd. correction	54.26	77.09	90.23	0.40
after 5th iteration	54.21	76.92	90.07	0.41
Automatic transcription				
HMM (1st stage)	36.69	61.26	80.17	0.82
HMM (2nd stage)	47.82	71.08	85.80	1.17
with bnd. correction	53.48	75.58	88.40	1.16
after 5th iteration	53.36	75.41	88.36	1.16

Table 3: Segmentation of TIMIT corpus. Percentage of correct boundaries and percentage of misaligned labels after the different segmentation stages.

The boundary correction performs better than the statistical correction proposed in [7]. However, the boundary refinement based on neural networks in [7] as well as the regression tree refinement investigated in [6] outperform our method in terms of total precision. This is not very surprising because statistical boundary corrections learn the particularities of the manual segmentation they are trained on while a text-independent correction method relies on properties of the signal only, which might disagree with the human labeler. Whether a higher agreement with the reference is desirable, eventually depends on what the segmentation is used for.

6. Conclusion

In this paper, we presented a phonetic segmentation method that is completely independent of manually segmented data, making it suitable for efficient segmentation of corpora in languages where no reference data is available. We showed that a boundary correction, that relies on signal properties only, can reduce the systematic errors made during HMM-based segmentation thus resulting in phone models better suited for the segmentation task. The boundary correction further allows to set very precise boundaries which is important for exact training of prosodic duration models.

7. References

[1] H. Romsdorfer, "Polyglot speech prosody control," in *Proceedings of Interspeech'09, Brighton*, 2009.

[2] G. Aversano *et al.*, "A new text independent method for phoneme segmentation," in *Proc. 44th MWSCAS*, 2001.

[3] L. Golipour and D. O'Shaughnessy, "A new approach for phoneme segmentation of speech signals," in *Proceedings of Interspeech'07*, 2007.

[4] F. Brugnara *et al.*, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[5] S. Pauws, Y. Kamp, and L. Willems, "A hierarchical method of automatic speech segmentation for synthesis applications," *Speech Communication*, vol. 19, 3, 1996.

[6] J. Adell *et al.*, "Comparative study of automatic phone segmentation methods for TTS," in *Proceedings of ICASSP'05*, 2005.

[7] D. Toledano, L. Gomez, and L. Grande, "Automatic phonetic segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625, 2003.

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

[9] S. Young, "The HTK Hidden Markov Model Toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[10] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.