Automatic Pitch-Synchronous Phonetic Segmentation with Context-Independent HMMs*

Jindřich Matoušek

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz

Abstract. This paper deals with an HMM-based automatic phonetic segmentation (APS) system. In particular, the use of a pitch-synchronous (PS) coding scheme within the context-independent (CI) HMM-based APS system is examined and compared to the "more traditional" pitch-asynchronous (PA) coding schemes for a given Czech male voice. For bootstrap-initialised CI-HMMs, exploited when some (manually) pre-segmented data are available, the proposed PS coding scheme performed best, especially in combination with CART-based refinement of the automatically segmented boundaries. For flat-start-initialised CI-HMMs, an inferior initialisation method used when no pre-segmented data are at disposal, standard PA coding schemes with longer parameterization shifts yielded better results. The results are also compared to the results obtained for APS systems with context-dependent (CD) HMMs. It was shown that, at least for the researched male voice, multiple-mixture CI-HMMs outperform CD-HMMs in the APS task.

1 Introduction

Automatic phonetic segmentation (APS) is a process of detecting boundaries between phones in speech signals. Since manual segmentation is labour-intensive and time-consuming, the automation of the process is very important especially when many speech signals are to be segmented. This is exactly the case of *unit selection*, a very popular and still the most prevalent *text-to-speech* (TTS) synthesis technique. Being a corpus-based concatenative speech synthesis method, the principle of unit selection is to concatenate pre-recorded speech segments (extracted from natural utterances in accordance with the automatically segmented boundaries) carefully selected from a large speech corpus according to phonetic and prosodic criteria imposed by the synthesised utterance. It is evident that automatic phonetic segmentation affects the quality of synthetic speech produced by a unit-selection-based TTS system.

The most successful approaches to the automatic phonetic segmentation are based on *hidden Markov models* (HMMs), a statistical framework adopted from the area of automatic speech recognition. There are many aspects that can affect the performance of the

* Support for this work was provided by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989, and by the Ministry of Education of the Czech Republic, project No. 2C06020.

V. Matoušek and P. Mautner (Eds.): TSD 2009, LNAI 5729, pp. 178–185, 2009.

HMM-based segmentation system such as the context dependency of HMMs, manner of initialisation of HMMs, training strategies, number of Gaussian mixtures used during modelling, speech coding schemes, etc. Various modifications and post-segmentation techniques were also proposed in order to increase the segmentation accuracy of the base APS system, see e.g. [1,2,3,4,5]. In this paper, the refinement based on *classification and regression trees* (CART) [3,5] is also applied when the segmentation accuracy of different coding schemes in a CI-HMM based APS system is evaluated.

In [6], *pitch-synchronous coding scheme* was proposed to enable more precise modelling of spectral properties of speech. It was shown that pitch-synchronous coding scheme outperformed the traditionally utilised pitch-asynchronous coding scheme when single-density context-dependent (CD) HMMs were employed. As some authors advocate for using context-independent (CI) HMMs with multiple Gaussian mixtures [2,5], the impact of different coding schemes on segmentation accuracy of APS system with CI-HMMs is examined in this paper.

The paper is organised as follows. CI-HMM-based APS system is briefly described in Section 2. In Section 3, various speech coding schemes are introduced. Experiments with different coding schemes and the results of the performance evaluation and their discussion are provided in Sections 4 and 5. Finally, conclusions are drawn in Section 6.

2 Automatic Phonetic Segmentation with CI-HMMs

The idea in APS is to apply similar procedures as for speech recognition. However, instead of the recognition, so-called *forced-alignment* is performed to find the best alignment between HMMs and the corresponding speech data, producing a set of boundaries which delimit speech segments belonging to each HMM. Briefly, each phone unit can be modelled by a context-dependent HMM (CD-HMM) or context-independent HMM (CI-HMM). In this paper, CI-HMMs are researched.

Firstly, the model parameters are to be initialised. Two initialisation strategies are usually employed [7]. When some (manually) segmented data are available, so called *bootstrap initialisation* can be applied utilising Baum-Welch algorithm with model boundaries fixed to the manually segmented ones (also called *isolated-unit training*). In this case, each HMM is initialised on its own phone-specific data. When no pre-segmented data are at disposal, so called *flat-start initialisation* is usually performed to set up all HMMs with the same data, typically corresponding to the global mean and variance.

Secondly, parameters of each model are trained on the basis of a collection of all speech data (described by *feature vectors*, often mel-frequency cepstral coefficients, MFCCs) with the corresponding phonetic transcripts. Typically, the *embedded training* strategy is employed in which models associated with the given phonetic transcript are concatenated and parameters of the composite model are simultaneously updated through the Baum-Welch algorithm. Simultaneously, in order to enable more precise modelling, the number of Gaussian mixtures can be incremented.

Finally, the trained HMMs are employed to align a speech signal along the associated phonetic transcript by means of Viterbi decoding. In this way, the best alignment

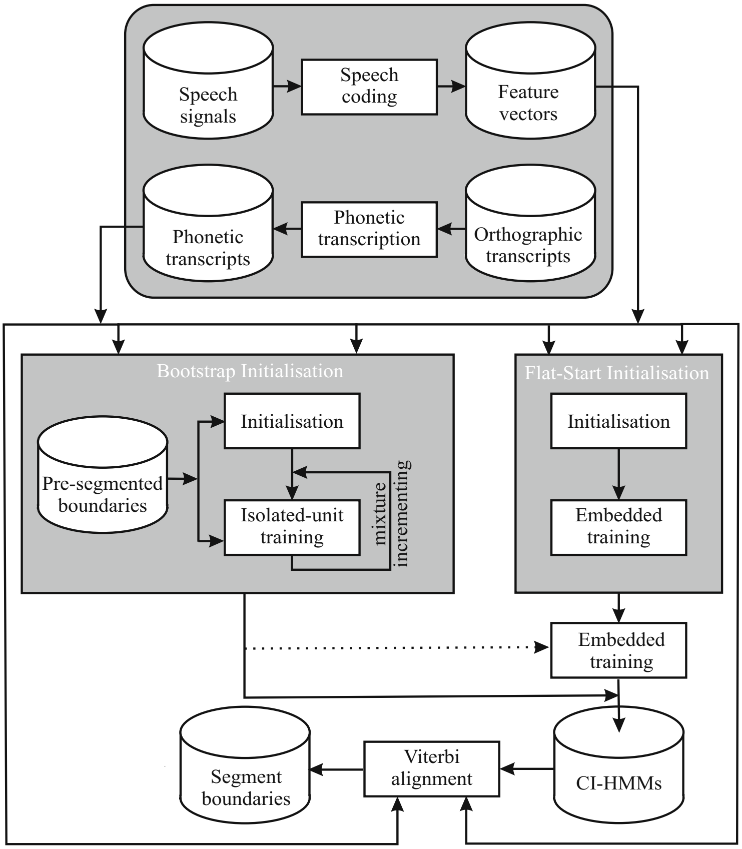

Fig. 1. Schematic view of a base automatic phonetic segmentation system with CI-HMMs, employing bootstrap or flat-start initialisation

between HMMs and the corresponding speech data is found, producing a set of boundaries which delimit speech segments belonging to each HMM. Thus, each phone-like unit is identified in the stream of speech signal and could be used for later purposes (e.g. in unit selection speech synthesis). A simplified scheme of a CI-HMM based APS system is given in Fig. 1.

Optionally, post-processing techniques can be employed to increase the segmentation accuracy by refining the initial segmentations from a base APS system. Statistically motivated approaches like classification and regression trees (CART), neural networks or support vector machines are often used to do the job. In this paper, CART-based refinement was performed similarly as in [3,5]; i.e. firstly, boundary-specific discrepancies between the automatically and manually segmented boundaries were learned by a CART, respecting phonetic type of the boundaries. Secondly, the automatic segmentations were refined by removing the boundary-specific biases according to the trained CART (see [6] for more details).

3 Speech Coding Schemes

The task of speech coding is to produce a sequence of feature vectors from speech signal of each utterance. The feature vectors are then used to train HMMs. It is obvious that the accuracy of boundary detection in an HMM-based APS system is limited by the manner the feature vectors are extracted from speech signals. Traditionally, a *pitch-asynchronous* (PA) coding scheme PA$\{l_{\mathrm{u}}/s_{\mathrm{u}}\}$ is employed for modelling speech. In this scheme, a uniform analysis frame of a given length l_{u} is defined and slid along the whole speech signal of an utterance with a fixed shift s_{u}. The length is usually set to comprise frequency characteristics of the speaker ($l_{\mathrm{u}} \approx 2T_0$ where T_0 is a maximum pitch period of the speaker). In automatic speech recognition, the shift is usually set to approx. 8-10 ms which roughly corresponds to T_0. In order to increase the segmentation resolution in APS, smaller shifts (4-6 ms) are also utilised.

In [6], *pitch-synchronous* (PS) coding scheme was proposed. In this scheme, each frame of speech to be extracted for coding is defined both by its position in a speech signal and its length. The positions and lengths of the frames are determined from *pitch-marks*, the locations of principal excitation of vocal tract (corresponding to glottal closure instants) in speech signals. In voiced speech, the position $p_{\mathrm{v}}^{(i)}$ of a frame $f^{(i)}$ ($i = 1, \ldots, N$) corresponds to a pitch-mark and the length $l_{\mathrm{v}}^{(i)}$ is set to $p_{\mathrm{v}}^{(i+1)} - p_{\mathrm{v}}^{(i-1)}$, which roughly corresponds to a double of the local pitch period (T_0). A very robust algorithm of pitch-mark detection was introduced in [8]. In unvoiced speech, no pitch-marks are defined because there is no activity of vocal cords during unvoiced speech regions. Therefore, standard PA coding scheme with a fixed frame length l_{u} and a fixed frame shift s_{u} is employed here. Smaller values such as $l_{\mathrm{u}} = $ 4–6 ms and $s_{\mathrm{u}} = $ 2–4 can be utilised. As a result, a sequence of frames $f^{(i)}$ consisting of the subsequences of

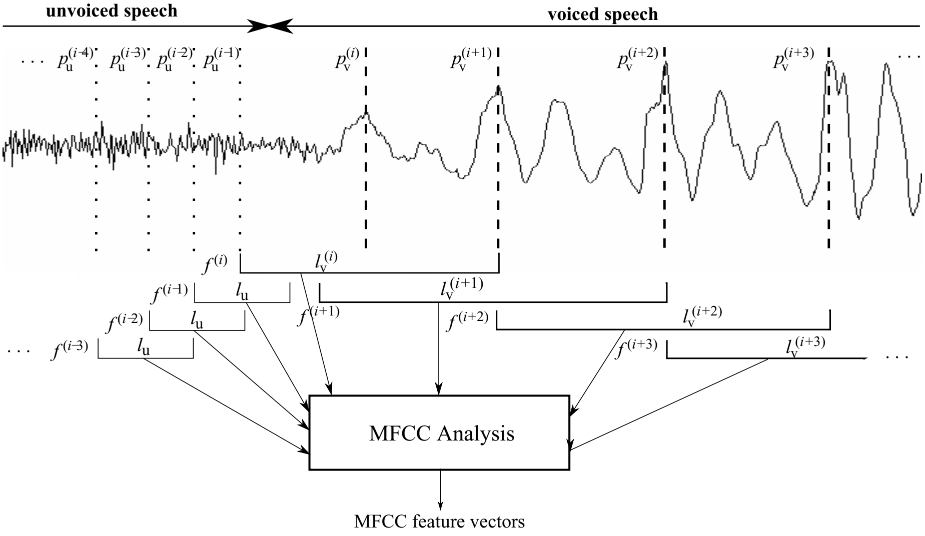
unvoiced speech
voiced speech
$\cdots$ $p_{\mathrm{u}}^{(i-4)}$ $p_{\mathrm{u}}^{(i-3)}$ $p_{\mathrm{u}}^{(i-2)}$ $p_{\mathrm{u}}^{(i-1)}$
$p_{\mathrm{v}}^{(i)}$
$p_{\mathrm{v}}^{(i+1)}$
$p_{\mathrm{v}}^{(i+2)}$
$p_{\mathrm{v}}^{(i+3)}$ $\cdots$
$f^{(i)}$ $l_{\mathrm{v}}^{(i)}$
$f^{(i-1)}$ l_{u}
$l_{\mathrm{v}}^{(i+1)}$
$f^{(i-2)}$ l_{u}
$f^{(i+1)}$
$f^{(i+2)}$ $l_{\mathrm{v}}^{(i+2)}$
$\cdots$ $f^{(i-3)}$ l_{u}
$f^{(i+3)}$ $l_{\mathrm{v}}^{(i+3)}$ $\cdots$
MFCC Analysis
MFCC feature vectors

Fig. 2. Illustration of a pitch-synchronous coding scheme

both voiced and unvoiced frames is available for coding. The illustration of the pitch-synchronous coding scheme is given in Fig. 2. More details about the pitch-synchronous coding scheme are given in [6].

4 Experiments and Results

In [6] two Czech phonetically and prosodically rich speech corpora (of a female and male voice) were researched. In this paper, only the male speech corpus is used, as the male voice is currently the main voice of the Czech TTS system ARTIC [9]. Hence, the results presented in this paper will be intentionally more specific for this speech corpus. The utterances included in the corpus were carefully selected, spoken by a professional speaker in an anechoic chamber, recorded at 16-bit precision with 48 kHz sampling frequency (later down-sampled to 16 kHz) and carefully annotated both on the orthographic and phonetic level [10]. Phonetic transcripts for all utterances plus some manual segmentations from a phonetic expert were available. In order to train the APS systems, a feature vector was computed for each frame according to various pitch-asynchronous coding schemes (the length of each frame was set according to our previous experiments to $l_{\mathrm{u}} = 25$ ms and shifts s_{u} were taken in the range 2-10 ms) and the pitch-synchronous scheme described in Section 3 using 12 MFCCs, log energy and their delta and delta-delta coefficients (39 coefficients for each frame in total). The corpus consists of 12,242 utterances (17.69 hours of speech excluding the pauses, 675,809 phone boundaries in total), 90 of them were segmented manually (11.71 minutes, 7,789 phone boundaries in total). 70 manually segmented utterances were used to initialise APS systems and 20 manually segmented utterances were used for testing. In order to reduce the labour-intensive and time-consuming manual segmentation, the amount of the manually segmented data was intentionally kept to minimum.

All experiments with the automatic phonetic segmentation were carried out following the scheme shown in Fig. 1 and using the HTK software [7]. Only experiments with different coding schemes were conducted – all other components of the APS system were fixed according to our previous experiments: each CI-HMM topology was fixed as 3-state left-to-right without any state skipping (with the exception of the pause models) with each state modelled using multiple Gaussian mixtures, the employ of both isolated (for initialisation) and embedded (for re-estimation) unit training procedures. CART-based refinement was performed using EST tool *wagon* [11].

The results of the automatic segmentation in terms of mean absolute error (MAE) and root mean square error (RMSE) for bootstrapped-initialised (BS) CI-HMMs (exploiting the pre-segmented data) and flat-start initialised (FS) CI-HMMs are shown in Table 1, or in Table 2, respectively. Let us note that, as gross segmentation errors are often to be avoided in unit selection speech synthesis, RMSE seems to be more suited for our comparisons. For BS CI-HMMs, the best results were obtained when no embedded training was performed at all; Viterbi alignment was performed right after the isolated-unit training. For FS APS systems no refinement was employed because no reference pre-segmented data are utilised for this kind of initialisation.

Table 1. Segmentation results for base and CART-refined BS APS systems. The number of mixtures was chosen according to the best segmentation results in terms of RMSE.

		Base		CART	
Coding scheme	# mixt.	MAE [ms]	RMSE [ms]	MAE [ms]	RMSE [ms]
PS	3	6.35	**10.02**	**5.35**	**8.66**
PA{25/10}	10	7.84	11.97	6.74	11.01
PA{25/8}	10	7.38	11.32	6.43	10.63
PA{25/6}	11	6.57	10.69	6.18	10.40
PA{25/4}	8	6.26	10.70	5.83	10.35
PA{25/2}	7	**5.90**	11.99	5.73	11.62

Table 2. Segmentation results for base FS APS systems. The number of mixtures was chosen according to the best segmentation results in terms of RMSE.

Coding scheme	# mixt.	MAE [ms]	RMSE [ms]
PS	2	11.88	18.41
PA{25/10}	1	9.99	17.32
PA{25/8}	1	**9.71**	**17.29**
PA{25/6}	2	10.13	17.71
PA{25/4}	5	10.94	18.35
PA{25/2}	16	13.16	20.84

5 Discussion

Looking at the results in Table 1 and Table 2, we can conclude, at least for the researched male voice:

- Pitch-synchronous coding scheme yields very good results when bootstrapped-initialised CI-HMMs are employed, especially after CART-based refining.
- PA{25/4} is the best one from pitch-asynchronous coding schemes and seems to be a good alternative to the pitch-synchronous coding scheme.
- PA{25/2} yields very good results for BS CI-HMMs in terms of MAE, but is prone to gross segmentation errors (see the RMSE score).
- As expected, BS CI-HMMs are much more superior to FS CI-HMMs. Therefore, one can conclude that it is worth preparing some manual segmentations to bootstrap the segmentation process.
- When no pre-segmented data are available (the case of flat-start initialisation), pitch-asynchronous coding schemes with longer parameterization shifts (8-10 ms) outperform the pitch-synchronous scheme.

Let us recall the segmentation accuracy of context-dependent HMMs for the same male voice, presented in [6]. The best results were achieved for the pitch-synchronous coding scheme, both for BS CD-HMMs (in case of the base system MAE was 8.11 ms, RMSE was 15.19 ms; after the CART-based refinement MAE decreased to 5.36 ms and RMSE to 12.79 ms) and for FS CD-HMMs (MAE was 9.62 ms and RMSE was

20.73 ms). Comparing both CD-HMM based APS and CI-HMM based APS, we can deduce that:

- CI-HMMs clearly outperform CD-HMMs when bootstrap initialisation is employed, both in terms of MAE and RMSE. Pitch-synchronous coding scheme seems to be the best choice in this case.
- As for flat-start initialisation, similar performance for CI-HMMs and CD-HMMs can be reported.

Considering the possibilities of post-segmentation corrections, it is obvious that there is more room for improvement for CD-HMMs. Indeed, analysing the effect of CART-based refinement, the greatest increase in segmentation accuracy was achieved for CD-HMMs (2.75 ms in MAE and 2.40 ms in RMSE; for CI-HMMs it was 1.00 ms in MAE and 1.36 ms in RMSE). As a result, the segmentation accuracy of CART-refined systems is almost the same for CD-HMMs and CI-HMMs in terms of MAE. As for RMSE, CART-refined CI-HMM based APS clearly outperforms CART-refined CD-HMM based APS.

6 Conclusion

In this paper, the use of the pitch-synchronous coding scheme within the context-independent HMM-based APS system was researched. The segmentation accuracy of such a system was compared to the segmentation accuracy of “more traditional” APS systems which utilise different pitch-asynchronous coding schemes. For bootstrap-initialised CI-HMMs, the proposed pitch-synchronous coding scheme performed best, especially in combination with CART-based refinement of the automatically segmented boundaries. For flat-start-initialised CI-HMMs, an inferior initialisation method used when no pre-segmented data are at disposal, standard pitch-asynchronous coding scheme with longer parameterization shifts yielded better results. The results were also compared to the results obtained for APS systems with context-dependent HMMs (as described in [6]). It can be shown that multiple-mixture CI-HMMs outperform CD-HMMs in the APS task, at least for the particular male voice under research.

The future work will be devoted to the further improvement of the performance of the APS system. Post-segmentation techniques, other than CART-based ones, are planned to be researched. Beside the segmentation accuracy itself, another important aspect of the APS system, the automatic detection of badly segmented boundaries (and especially gross segmentation errors) will be researched. This aspect is very important from the point of view of unit selection speech synthesis, because the badly segmented units (which cause degradation of the resulting speech) can be removed from unit inventories and thus they can be ignored during speech synthesis.

References

1. Matoušek, J., Tihelka, D., Romportl, J.: Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction. In: Proceedings of Interspeech, Geneve, Switzerland, pp. 301–304 (2003)

2. Toledano, D., Gómez, L., Grande, L.: Automatic Phonetic Segmentation. IEEE Transactions on Speech and Audio Processing 11(6), 617–625 (2003)
3. Adell, J., Bonafonte, A.: Towards Phone Segmentation for Concatenative Speech Synthesis. In: Proceedings of Speech Synthesis Workshop, Pittsburgh, U.S.A, pp. 139–144 (2004)
4. Lee, K.S.: MLP-Based Phone Boundary Refining for a TTS Database. IEEE Transactions on Audio, Speech and Language Processing 14(3), 981–989 (2006)
5. Park, S.S., Kim, N.S.: On Using Multiple Models for Automatic Speech Segmentation. IEEE Transactions on Audio, Speech and Language Processing 15(8), 2202–2212 (2007)
6. Matoušek, J., Romportl, J.: Automatic Pitch-Synchronous Phonetic Segmentation. In: Proceedings of Interspeech, Brisbane, Australia, pp. 1626–1629 (2008)
7. Young, S., et al.: The HTK Book (for HTK Version 3.4). Cambridge University, Cambridge (2006)
8. Legát, M., Matoušek, J., Tihelka, D.: A Robust Multi-Phase Pitch-Mark Detection Algorithm. In: Proceedings of Interspeech, Antwerp, Belgium, pp. 1641–1644 (2007)
9. Matoušek, J., Tihelka, D., Romportl, J.: Current state of czech text-to-speech system ARTIC. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 439–446. Springer, Heidelberg (2006)
10. Matoušek, J., Tihelka, D., Romportl, J.: Building of a Speech Corpus Optimised for Unit Selection TTS Synthesis. In: Proceedings of International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco (2008)
11. Taylor, P., Caley, R., Black, A., King, S.: Edinburgh Speech Tools Library: System Documentation (1999), `http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/`