Automatic Phonetic Segmentation for a Speech Corpus of Hebrew

Nikša Jakovljević, Dragiša Mišković, Darko Pekar, Milan Sečujski, Vlado Delić
Department of Power, Electronics and Communication Engineering
Faculty of Technical Sciences
Novi Sad, Serbia
jakovnik@uns.ac.rs, dragisa@uns.ac.rs, darko.pekar@alfanum.co.rs, secujski@uns.ac.rs, vdelic@uns.ac.rs

Abstract—**This paper presents our study on different phonetic segmentation methods based on hidden Markov models evaluated against a Hebrew speech corpus. We investigated methods for fully automatic phonetic segmentation using only the corpus which should be segmented and automatically generated phonetic transcriptions. A new method for phonetic boundary correction based on spectral variation of the speech signal is proposed. The proposed method increased the boundary correctness of the baseline HMM segmentation system from 30.2%, 59.5% and 86.2% of automatic boundary marks with error smaller than 5, 10 and 20 ms respectively, to 52.3%, 76.3% and 90.7%.**

Key words – automatic phonetic segmentation; boundary correction;

I. INTRODUCTION

Contemporary text-to-speech (TTS) systems based on segment concatenation require large speech databases segmented into phones in order to produce high-quality speech. The quality of a TTS system is dependent on the accuracy with which the speech corpus is segmented and labeled as well. Inaccurate time stamps and incorrect labels can result in noticeable errors in the synthetic speech. Traditionally, accurate segmentation is performed by human experts in the field of acoustic phonetics. Manual segmentation is an extremely time-consuming activity (it requires over two hours for a one minute recording [1]) and it is difficult to keep boundary placement consistent especially when more than one labeler is involved [2]. In case of large corpora, which are nowadays common, automatic segmentation is necessary.

Besides TTS, many areas in speech processing require algorithms that perform time alignment of a speech signal with phonetic transcriptions, and therefore many such algorithms have been developed. These algorithms can be classified into 2 broad classes according to whether they use prior knowledge about text content for the initial segmentation of a speech file [1].

Linguistically unconstrained algorithms use a bottom-up strategy which exploits only acoustic information contained in a speech signal in order to detect phone transitions. These algorithms do not depend on phonetic transcriptions, and thus are suitable for multilingual application or low bit rate speech coding. It is suggested to classify them into two broad categories, i.e. model free and model based algorithms [3]. Model free algorithms define a change function that directly measures the spectral variation of the acoustic signal and use it as a transition penalty [4], [5]. On the other hand, model based algorithms assume that potential segmentation points correspond to sequential model changes, i.e. for each frame two statistical hypotheses are evaluated: H_0 – that the frame belongs to local model and H_1 – that the frame is a transition point [1], [6], [7]. This approach is used for speaker segmentation in [8], [9] as well.

Linguistically constrained algorithms exploit both the speech signal and the information about its content to generate a reliable segmentation. They are based on generative speech models such as dynamic time warping (DTW) and hidden Markov models (HMM) as well as discriminative models such as artificial neural networks (ANN) [10]-[20]. These algorithms are more accurate than linguistically unconstrained algorithms, but they require training data and they are speaker and language dependent.

A common way to evaluate the accuracy of segmentation is by comparing the resulting segmentation to manual segmentations and includes the calculation of figures of merit such as mean error, root mean square error, and the percentage of errors smaller than a tolerance value. The last is the most frequent and a typical value for tolerance is 20 ms. Besides these, so called direct figures of merit, there are indirect figures of merit, which evaluate word error rate of the recognizer used in segmentation stage or the subjective quality of speech synthesizer [11]. There is also a type of error which is specific for linguistically unconstrained algorithms, which represents the portion of points which are incorrectly identified as phone boundaries – so called false alarms.

The performance of linguistically unconstrained algorithms is about 76 % hit rate with the tolerance of 20 ms and limiting over-segmentation to a minimum [1]. The hit rate for algorithms based on HMM is 85-90% [11], [16] for the same tolerance, and with additional boundary correction this rate can attain up to 96% [10], [11]. In cases when manually segmented speech material is not available for model training hit rate is about 90% (with additional boundary corrections) [16]. For the sake of comparison, the discrepancy between labels segmented manually by different labelers is 97% for the same tolerance [11].

This study investigated fully automatic phonetic segmentation, i.e. the case when only the corpora to be segmented and the automatically generated transcriptions are

This research was supported in part by the Ministry of Education and Science of the Republic of Serbia grant number TR32035.

used. We have chosen an approach similar to the one presented in [16], which is based on HMM with additional boundary correction using a distance between feature vectors. In comparison to the methods presented in [10]-[15], no manually annotated data were used, making the analyzed approach language independent.

The paper is organized as follows. Section II describes the speech corpus of Hebrew which is used. In section III, an overview of the segmentation procedure is given. Results are discussed in section IV, which is followed by conclusions and directions for further research.

II. SPEECH CORPUS

The corpus of Hebrew consists of about 900 sentences spoken by one male professional speaker, sampled at 44.1 kHz. For the purpose of automatic segmentation speech signal was down-sampled to 22 kHz. The automatic phone transcription of the corpus largely corresponds with the exact content of speech signal, but there are differences caused by different pronunciation variants or rapid speech phenomena like assimilation and elision.

For the test set, 50 sentences were chosen randomly, and annotated manually. During the manual segmentation of the test set only label boundaries were corrected, but phone transcriptions remained the same even if they contained errors. In this way evaluated performance on the test set is closer to the performance on the rest of the corpus which contains erroneous transcriptions too. If the phone was not pronounced (usually glottal stops and the voiceless glottal fricative) its start position is set to be the start position of the following phone, furthermore if the phone was substituted by another phone, it was treated as the correct one. All these erroneous phones are marked as such in order to have special treatment in evaluation. We evaluate the performance as hit rates for given thresholds using (1) all phones and (2) only correct phones, in order to estimate the influence of wrong phonetic transcriptions on the procedure of segmentation.

III. SEGMENTATION

Segmentation algorithms based on HMMs consist of two phases: *i*) force alignment and *ii*) boundary correction. In the force alignment phase the Viterbi algorithm uses HMMs and automatic transcriptions to obtain rough phonetic boundaries. Although HMMs are nowadays the dominant approach in speech recognition, they do not produce precise phonetic boundaries because HMM objective function used in training phase is chosen to indentify phonetic segments, not to produce precise phonetic boundaries [11]. Moreover, the speech feature extraction mechanism has limited resolution because the features are extracted over 20-30 ms windows every 5-10 ms. The standard method to train HMMs for recognition is embedded training, however in this way phonetic boundaries can be changed, thus for phonetic segmentation purposes better results can be obtained by isolated phoneme training [16]. The common modelling unit in speech recognition is a context-dependent phone, but in phonetic segmentation context-independent models are more robust in case of non-stationary phones (e.g. affricates) [11].

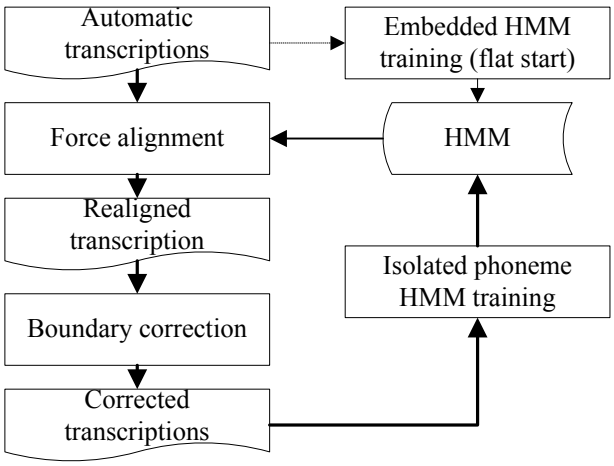

Figure 1. Block diagram of automatic phonetic segmentation procedure

As noted above, phonetic boundaries set by HMMs are not at transition points, but very close to them (about 85-90% of them are within the 20 ms range), and thus they can be corrected using: additional phone transition models based on GMM [12], ANN [19], [20] or information about spectral changes [11], [17]. On the other hand, these errors are systematic, and thus it is possible to statistically model them and use the resulting model to correct them [10], [11]. In the case when manually segmented data is not available, the only possible approach is to use the information about spectral changes, because it does not require model training data.

The block diagram of the procedure of automatic phonetic segmentation explored in this study is shown in Fig. 1. First, HMMs are trained with so called flat start procedure [21] which does not require initial phonetic boundaries. These HMMs are then used for alignment of phonetic transcriptions with audio material using Viterbi algorithm (force alignment) [21]. These relabelled data are used as input for the boundary correction procedure which exploits spectral changes to set phonetic boundaries. In this study we implemented procedures described in [11] and [4]. The corrected transcriptions are then used for isolated phoneme training of HMMs because this approach results in HMMs which are more accurate than the HMMs trained with flat start [10], [16]. These new models are then used in the force alignment step of the next segmentation cycle. This segmentation cycle can be repeated if required.

A. *Boundary correction*

In the first approach presented in [16] it is assumed that phone boundaries can be considered as a transition phase between relatively stable centres of the phones, and that they can be detected measuring the distance to the phone central frames. Given two adjacent phones P_i and P_{i+1}, with indexes of central frames c_i and c_{i+1}, the boundary b is given by:

$$b = \left\lfloor \frac{b_i + b_{i+1}}{2} \right\rfloor \quad (1)$$

$$b_i = b \mid D(\mathbf{x}_{c_i}, \mathbf{x}_b) \geq D(\mathbf{x}_{c_{i+1}}, \mathbf{x}_b) \wedge \forall c_i < j < b \; D(\mathbf{x}_{c_i}, \mathbf{x}_j) < D(\mathbf{x}_{c_{i+1}}, \mathbf{x}_j) \quad (2)$$

$$b_{i+1} = b \mid D(\mathbf{x}_{c_i}, \mathbf{x}_b) \leq D(\mathbf{x}_{c_{i+1}}, \mathbf{x}_b) \wedge \forall b < j < c_{i+1} \; D(\mathbf{x}_{c_i}, \mathbf{x}_j) > D(\mathbf{x}_{c_{i+1}}, \mathbf{x}_j) \quad (3)$$

where $D(\cdot,\cdot)$ represents the Euclidean distance between two frames and $\mathbf{x}_j$ is the frame with the time index j.

The second approach is based on the detection of local maxima in the delta cepstral change function *DCF* [4] defined by:

$$DCF(j) = \frac{\sum_{f=1}^{F} \mathbf{d}_j(f)}{\max_i \sum_{f=1}^{F} \mathbf{d}_i(f)} \quad (4)$$

where j is a frame index, F is the number of features in a feature vector (frame) and $\mathbf{d}_j(f)$ is the absolute cepstrum slope estimated for j^{th} frame and f^{th} feature defined by:

$$\mathbf{d}_j(f) = \frac{|\mathbf{x}_{j+o}(f) - \mathbf{x}_{j-o}(f)|}{\max_i |\mathbf{x}_{i+o}(f) - \mathbf{x}_{i-o}(f)|} \quad (5)$$

where o is a predefined offset.

IV. Experiments and Results

A. The baseline system

Our baseline speech segmentation system uses HMMs to align phonetic labels to speech signals. We adapted our automatic speech recognition system [22] to perform phonetic segmentation. Each phoneme is mapped to one context dependent phone model (triphone) with the exception of plosives and affricates, where occlusion and explosion of plosives and occlusion and friction of affricates are separate modelling units. This approach requires special treatment of context for the occlusions (explosions and frictions of affricates) because they have the same right (left) context, thus as a right (left) context use successive (preceding) phoneme.

The number of HMM states is proportional to the duration of the phone which is modelled and varies from 2 for explosions of plosives up to 12 for long stressed vowels. Such a high number of states is motivated by results in [13] which suggest that models with higher number of states give better alignment precision. Although in the published experiments all models have the same number of states, in this study the number of states is proportional to the phone duration in order to model phone dynamics better [12]. The state emitting distribution is modelled by a single multivariate normal (Gaussian) distribution.

The frame size is 20 ms and the frame step is 2 ms in the training phase and 5 ms in the testing phase. The frame step in the test phase is smaller than in the testing phase so as to efficiently estimate model parameters, but it requires some additional changes in the extraction procedure as in [23]. An effect of using a 2 ms frame shift in the testing phase is the extension of segmentation process, but boundary correctness remains the same. For each frame a 26 dimensional vector composed of 12 MFCCs and normalized energy, as well as their first order derivatives is calculated.

In force alignment procedure short pauses, glottal stops and the voiceless glottal fricative are optional because they can be omitted during pronunciation. Even if, during alignment, they are identified as non-existing, they remain in the transcription but with zero duration.

B. Plosive and affricate splitting

If the level of noise is sufficiently low, the border between occlusion and explosion or occlusion and friction in case of voiceless plosives and affricates respectively can be detected by a sudden rise in signal energy. On the other hand, in case of voiced plosives and affricates this boundary is manifested by a sudden rise of energy at high frequencies (over 600 Hz). This rise of energy at high frequencies is exploited for plosive and affricate splitting.

To remove low frequency components the audio signal is filtered by a high pass FIR filter with constant group delay (pass band frequency is 1 kHz, stop band frequency is 600 Hz, maximum attenuation in pass band is 0.5 dB and minimum attenuation in stop band is 40 dB). Using a filter with constant group delay was necessary for the purpose of synchronisation with the original audio file. The energy of the filtered signal is computed every 4 ms over a 20 ms window.

This procedure is very accurate i.e. 90% of borders between occlusion and explosion and occlusion and friction are within a 5 ms range from their true positions. The most common errors are the consequence of coarticulation with neighbouring phones which leads to the existence of bursts of noise during the occlusion.

C. Discussion

In case of boundary correction based on the distance from central frames, the frame window duration was varied from 10 ms up to 30 ms, and the best performance was obtained with 10 ms. As features we used only spectral features i.e. 12 MFCCs and normalized energy, the spectral features and their first derivatives, as well as the spectral features and their first and second derivatives. The best result was obtained with normalized spectral features (the features are scaled so that in a single file variance per each feature is 1). The results (see table I) are slightly inferior to those given in [16]. A detailed analysis showed that this boundary correction is inefficient in case of non-stationary phones such as glottal stops, sequences of similar phones with significant coarticulation, as well as in case of significant misalignment errors. This was the reason why we tried to detect spectral changes.

The correction based on spectral change used only spectral features i.e. 12 MFCCs and normalized energy computed every 2 ms. The window duration was varied from 10 ms up to 30 ms but the best results were obtained with 20 ms. The predefined offset for *DCF* is 10 ms. Since *DCF* was calculated every 2 ms, instead of the maximum of *DCF* as in [4], region with maximum was considered as well as the initial phone boundary. The results are shown in table II.

Although more accurate HMMs trained on corrected boundaries result in better segmentation, additional boundary correction does not improve these boundaries significantly.

TABLE I. SEGMENTATION PERFORMANCE FOR VARIOUS METHODS. HIT RATES OR PERCENTAGE OF ERRORS SMALLER THAN GIVEN TOLERANCE VALUES EVALUATED TAKING IN ACCOUNT ALL LABELS (ALL) AND ONLY CORRECT LABELS (CORR).

Segmentation method		**Tolerance**		
		5 ms	**10 ms**	**20 ms**
Baseline (HMM only)	all	28.5	58.2	85.4
	corr.	30.2	59.5	86.2
Correction based on distance	all	45.2	70.8	89.2
	corr.	46.0	71.8	90.1
Correction based on *DCF*	all	51.3	75.1	89.9
	corr.	52.4	76.3	90.7
Correction based on distance (iter.)	all	44.7	70.0	88.8
	corr.	45.3	70.9	89.1
Correction based on *DCF* (iter.)	all	51.5	74.9	89.8
	corr.	52.5	75.8	90.6

V. CONCLUSION AND FURTHER DIRECTIONS

In this paper several segmentation methods completely independent of manually segmented data were presented. These methods proved to be suitable for phonetic segmentation of corpora in languages where no manually labelled data is available. Experiments showed that boundary correction based on signal spectrum can significantly reduce systematic error. In comparison to the method based on *DCF*, the method based on distance shows slightly inferior performance. The last method makes significant errors in case of non-stationary phones (glottal occlusions) and in case of sequences of similar phones pronounced relatively fast. Contrary to our expectation, an additional iteration with corrected labels resulted in slightly inferior performance, and we assume that the reason for this are incorrect transcriptions. Since phonetic transcriptions contain incorrect labels, further directions should be the development of a method which detects and corrects such errors automatically.

ACKNOWLEDGMENT

The authors would like to thank Ron Hasson of Aharon Group, Israel, who provided us with audio files and their initial phonetic transcriptions and provided finance support for this research.

REFERENCES

[1] G. Almpanidis, M. Kotti and C. Kotropouls, "Robust Detection of Phone Boundaries Using Model Selection with Few Observations," in IEEE Trans. Audio Speech and Lang. Process. vol 17. pp. 287-298, Feb. 2009.

[2] A. Ljolje, J. Hirschberg, and J. P. van Santen, "Automatic speech segmentation for concatenative inventory selection," SSW2-1994, USA, pp. 93-96, 1994.

[3] A. Esposito and G. Aversano, "Text Independent Methods for Speech Segmentation," Nonliear Speech Modeling, LNAI 3445, pp. 261-290, 2005.

[4] C. Mitchell, M. Harper and L. Jamieson, "Using explicit segmentation to improve HMM phone recoginition," in IEEE Int. Conf. Acoust. Speech Signal Process 1995., vol 1. pp. 229-232, 1995.

[5] F. Brugnara, R. de Mori, D. Guillani and M. Omologo, "Improved connected digit recognition using spectral variation functions," in Int. Conf. Spoken Lang. Process. 1992., vol. 1, pp. 627-630, 1992.

[6] G. Flammia, P. Dalsgaard, O. Andersen and B. Lindberg, "Segment Based Variable Frame Rate Speech Analysis and Recognition Using Spectral Variation Function," in Int. Conf. Spoken Lang. Procss 1992, pp. 983-986, 1992.

[7] J. C. Segura-Luna, J. M. Soler, A. M. Peinado, V. Sanchez and A. Rubio, "Signal Segmentation into Spectral Homogeneous Units" in Eurospeech '90, pp. 1251-1254, 1990.

[8] P. Delacourt and C. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," Speech Commun., vol. 32, pp. 111-126, 2000.

[9] M. Cettolo, M. Vescovi and R. Rizzi, "Evaluation of BIC-based algorithms for audio segmentation," Comput. Speech, Lang, vol 19. pp. 147-170, 2005.

[10] J. Matoušek, D. Tihelka and J. Psutka, "Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction," in Eurospeech '03, pp. 301-304, 2003.

[11] D. T. Toledano and L. A. Hernandez Gomez, "Automatic Phonetic Segmentation," in IEEE Trans. Audio Speech and Lang. Process. vol 11. pp. 617-625, .Nov. 2003.

[12] L. Wang, Y. Zhao, M. Chu, J. Zhou and Z. Cao, "Refining Segmental Boundaries for TTS Database Using Fine Contextual Dependent Boundary Models," in IEEE Int. Conf. Acoustic, Speech Signal Process. 2004, pp. 641-644, 2004.

[13] J. Dines, S. Sridharan and M. Moody, "Automatic Speech Segmentation with HMM," in 9th Int. Conf. Speech Science & Technology, pp. 544-549, 2002.

[14] J. A. Gomez and M. Calvo, "Improvements on Automatic Speech Segmentation at the Phonetic Level," in CIARP 2011, pp 557-564, 2011.

[15] J. Adell, A. Bonafonte, J. A. Gomez and M. J. Castro, "Comparative study of Automatic Phone Segmentation Methods for TTS," in Int. Conf. Acoustic, Speech Signal Process. 2005, pp.319-312, 2005.

[16] S. Hoffmann and B. Pfister, "Fully Automatic Segmentation for Prosodic Speech Corpora," in Interspeech '10, pp. 1389-1392, 2010.

[17] Y. J. Kim and A. Conkie, "Automatic segmentation combining an HMM-based approach and spectral boundary correction," in 7th Int. Conf. Spoken Language Process. 2002, pp 145-148, 2002.

[18] M. Sharma and R. Mammone, "Automatic Speech Segmentation Using Neural Tree Networks," in IEEE Workshop on Neural Net. for Signal Process., pp. 282-290, 1995.

[19] D. T. Toledano, "Neural Network Boundary Refining for Automatic Speech Segmentation," in IEEE Int. Conf. Acoustic, Speech Signal Process. 2000, pp. 3438-3441, 2000.

[20] K. S. Lee, "MLP Base Phone Boundary Refining for a TTS Database," in IEEE Trans. Audio Speech and Lang. Process. vol. 14, pp. 981-989, 2006.

[21] S. Young *et al.*, The HTK Book, version 3.4, Cambridge University Press, 2006.

[22] V. Delić, M. Sečujski, N. Jakovljević, M. Janev, R. Obradović and D. Pekar, "Speech Technologies for Serbian and Kindred South Slavic Languages," 9th Chap. in the book: Advances in Speech Recognition, N. R. Shabtai (Ed.), 2010.

[23] D. Pekar, N. Jakovljević, M. Janev, D. Mišković and V. Delić, "On the Use of Higher Frame Rate in the Training Phase of ASR," 14th WSEAS Int. Conf. on Latest Trends on Comp., pp. 127-130, 2010.