VOICE SEGMENTATION SYSTEM BASED ON ENERGY ESTIMATION

Raissa B. Rocha, Virginio V. Freire and Marcelo S. Alencar

Federal University of Campina Grande (UFCG)
Federal University of Sergipe (UFS)
Institute of Advanced Studies in Communication (Iecom)

ABSTRACT

Voice segmentation is used in speech recognition and system synthesis, as well as in phonetic voice encoders. This paper describes an implicit speech segmentation system, which aims to estimate the boundaries between phonemes in a locution. To find the segmentation marks, the proposed method initially locates reference borders between silent periods and phonemes, and vice versa measuring energy in short duration periods. The phonetic boundaries are found by means of energy encoding in the region delimited by the reference marks, which were initially detected. To evaluate the performance of the proposed system, an objective evaluation using 50 locutions was performed. The system detected 72.41% of the segmentation marks, in which, 77.6% were detected with an error less or equal to 10 ms and 22.4% of the boundaries were found with an error between 10 and 20 ms.

Index Terms— Voice segmentation, energy detection, objective evaluation.

1. INTRODUCTION

The speech, that represents a determined message to be transmitted to a listener, is formed by the junction of short sounds called phones. To develop an application for voice signal processing it is necessary that the speech be segmented to improve the robustness of the algorithm for voice recognition. The improvement of segmentation techniques provides increases the man-machine interaction by the recognition of voice commands by the machine or generation of synthetic speech [1].

A speech segmentation system aims to determine the boundaries that separate the essential elements, such as words, syllables or phonemes, in a locution [2]. This system can be used in phonetic speech coders, as well as in automatic speech synthesis and recognition systems.

In speech recognition, which has the objective of recognizing each word or sentence pronounced by the speaker, the segmentation becomes important to build a database of segmented voice. This database is used in the training phase of models related to the phonetic subunits in systems that make use of statistical modeling, such as the Hidden Markov Models (HMM). Otherwise, in the training phase, the phrases are segmented uniformly to generate the first estimates which will be refined during the next training iterations [3, 4].

The speech segmentation is also used in voice synthesis or Text To Speech (TTS) conversion, which consists of a change of domain for the representation of information, from the written form to the spoken one. In these systems, there is initially an analysis of the text, which results in the standardization and phonetic transcription. Then, a voice signal synthesis is made, which may be done, for example, with the selection and concatenation of the acoustic units present in a speech database formed by the segmentation of voice signals.

Phonetic encoders produce voice signal with low bit rate. To achieve this objective, the speech signal, applied to the input of the encoder emitter, must be phonetically segmented in order to extract each phoneme prosodic information, such as duration, energy and pitch, and send this information to the receiver to perform the voice signal synthesis. Thus, the speech segmentation system directly affects the performance of the phonetic encoders, since a robust segmentation is necessary for the exact extraction of the parameters and for the correct concatenation synthesis in the receiver [5].

The literature classifies the speech segmentation systems according to the presence or absence of a linguistic category and the existence of acoustic observations.

It is understood by linguistic category information the phonetic transcription of a phrase, that may or may not be input to the segmentation system. The acoustic observations consist on information extracted from the speech signal, normally represented by a parameter vector with information of the speech signal assigned to short period windows.

Speech segmentation systems can be classified as implicit or explicit. The implicit segmentation happens when the linguistic category is not considered in the segmentation process, and only the acoustic observations for the system are used to generate the segmentation boundaries. The explicit segmentation utilizes the phonetic transcription (linguistic information) to generate the segmentation marks. Therefore, in this

The authors would like thanks to Federal University of Campina Grande, Federal University of Sergipe and Institute of Advanced Studies in Communication .

type of segmentation, phonetic transcriptions of speech to be segmented must be generated in advance and used as input to the segmentation system.

The voice segmentation can be automatic or manual. In the first case, the boundaries of the acoustic units are obtained by computational techniques that highlight the boundaries from acoustic characteristic. An specialist is needed in the manual segmentation, with tools or software that are capable of displaying the voice waveform, the spectrogram and possibly another acoustic information. This procedure takes a considerable amount of time, and is a tedious task. However, the manual segmentation is important to evaluate an automatic system.

This paper discuss the development of a voice segmentation system that uses energy observation of the locutions. Initially, reference boundaries are obtained by recording the difference between the energy during the silence and voiced segments. After that, the segmentation marks are obtained by encoding the energy of the locution of the reference boundaries previously obtained.

This paper has three additional sections. Section 2 describes the necessary steps to implement the voice segmentation system. Section 3 presents the results obtained and Section 4 presents the conclusions and future works.

2. VOICE SEGMENTATION SYSTEM

The speech segmentation system to be enhanced is classified as implicit, because it does not use phonetic transcription data to segment the speech, and provide the initial and final instants of each phoneme in a phrase.

Some studies found in the literature on speech segmentation use statistical methods, such as Hidden Markov Models, and refining systems, in which prior knowledge of the phonetic transcription from a phrase is necessary, among other features [2, 3, 6–13]. The system under study segments the speech into phonemes by the observation of a prosodic feature of speech signal, the pitch.

2.1. Description of the Segmentation System Improvements

The speech segmentation method has two steps, which are described in the following.

2.1.1. 1st Step: Identification of audible regions

The first stage consists of making an initial segmentation with the identification of audible and non-audible regions using short duration energy. This parameter presents significantly high values for audible regions in a phrase, which makes it possible to distinguish silent from voiced intervals.

This procedure improves the performance of the segmentation by obtaining the boundaries between phonemes and silence, and vice-versa, witch are easier to identify.

According to the literature, the short duration energy is calculated using windows whose size vary between 20 samples for a high-pitched voice and 250 samples to a low-pitched voice. In practice, for a sampling rate of about 10 k samples/s, a window between 100 and 200 samples ($10\text{ ms} < t < 20\text{ ms}$) must be used.

The sampling rate of the phrases used in the development of the segmentation system is 22,050 samples/s. Thus, the energy calculation is made with a window of 500 samples, using a window shift of 20 samples. Figure 1 illustrates an example of reference boundaries acquisition, located between phonemes and silence, and vice-versa.

Fig. 1. Example of how to obtain borders between silence and phonemes, and vice versa.

2.1.2. 2nd Step: Identification of New Reference Boundaries

The second step consists of a new segmentation made in the audible regions, delimited by reference boundaries, with the goal of finding new reference boundaries present in each audible region.

The identification of new reference borders is performed by the voice signal energy. The energy is calculated for each interval of predetermined duration (200 samples). Figure 2.2 illustrates the behavior of the energy (red) in relation to a phrase (blue).

According to Figure 2, it is observed that the new reference borders are located at the beginning and at the end of valleys of the energy curve, because most of the boundaries between phonemes are in regions in which the energy is increasing or decreasing.

To identify the initial and final instants of each valley, an encoding of the energy is performed. Initially, the phrase to be segmented is divided into four regions. For each region an average energy is obtained. This parameter is used as a threshold and the code 1 is assigned to the energy values above it. For energy values below the threshold, the code is 0.

As a result of this procedure, a matrix formed by regions of zeros and ones is obtained. To identify the reference boundaries, a search is performed between the transition regions of ones and zeros. Thus, new reference borders are found, since they are located in the transition region of the associate array.

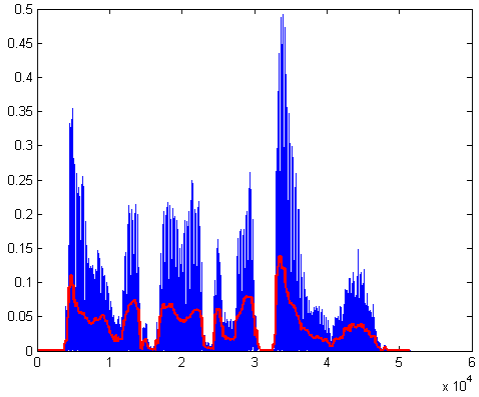

Fig. 2. Energy curve related to the phrase.

3. RESULTS

To evaluate the performance of the proposed segmentation system, an objective evaluation was performed. Hence, the value of the boundaries obtained by the segmentation system were compared to results obtained manually. The database used in the tests was manually segmented [2] [3] to be use in the research. It is composed of 200 phrases, recorded in the city of So Paulo. The sentences were recorded with a rate of 22.05 k samples/s and quantized with 16 bits by sample. The locutions have, on average, three seconds and were recorded with the lowest possible noise.

For the evaluation the phrases were select from database 50 locutions. The Table 1 presents the results of segmentation rate, the number of false boundaries obtained, as well as the quantity of boundaries not detected in each sentence.

According to Table 3.1, the proposed segmentation system was able to detect 72,41% of the border segmentations, and 77,6% of the segmentation marks presented an error lower than 10 ms (221 samples), while 22,4% of the boundaries obtained had errors between 10 and 20 ms (221 to 442 samples). Furthermore, the developed system did not detect, on average, 7,6 boundaries, and located, on average, 6,74 false boundaries.

4. CONCLUSIONS AND FUTURE WORKS

Digital processing techniques are present in several applications, such as voice mail, automatic voice systems, biometric identification, language identification, voice dialing, residential automation. Voice commands are used, as well as reading systems for the blind. Those applications show the unequivocal contributions of vocal communications between man and machine, and they include the voice recognition and text to speech conversion systems, for which the segmentation system performs and important task.

This paper describes a speech segmentation system based on observation of a prosodic characteristic of the voice signal, the energy.

To find the segmentation marks, the proposed method initially located the reference frontiers. Then, new reference boundaries are found using an energy encoder that operates in the locution region defined by reference boundaries initially detected.

The proposed segmentation system was tested using 50 segmentation locutions. To verify its robustness, the obtained results were compared which segmentation marks achieved using manual segmentation. In general, the system is capable to detect 72,41% of the segmentations marks, in which 77,6% were detected with an error smaller than 10 ms, and 22,4% had and error between 10 and 20 ms, when compared with boundaries resulting from manual segmentation. Furthermore, the system presented, on average, 6,74 false frontiers and 7,6 boundaries that were not detected.

The proposed segmentation algorithm presents a low complexity. The development does not depend on a robust database to train probabilistic models, as in the case HMM segmentation. Moreover, no prior knowledge of phonetic transcription is necessary for the segmentation. The proposed system has competitive results as compared to usual systems, without the use of a refinement of the obtained results.

As a future work, the authors intend to develop a refinement method to be used in the proposed system, to decrease the segmentation errors and the eliminate false boundaries.

REFERENCES

[1] Hamed Talea and Khashayar Yaghmaie, "Automatic Visual Speech Segmentation," *Communitations Software and Networks (ICCSN)*, 2011.

[2] A. M. Selmini, "Sistema Baseado em Regras para o Refinamento da Segmentao Automtica de Fala.," Tese de doutorado, Universidade Estadual de Campinas, Campinas, Brasil, Agosto de 2008.

[3] E. D. S. Paranagu, "Segmentao automtica do sinal de voz para sistemas de converso texto-fala," Tese de doutorado, Universidade Federal do Rio de Janeiro, Maro 2012.

[4] S.Harish, P.Vijayalakshmi, and T. Nagarajan, "Significance of Segmentation in Phoneme Based Tamil Speech Recognition System ," 2011.

[5] R. B. Rocha, "Desenvolvimento de um Codificador de Voz Pessoal de Baixa Taxa Baseado em Modelos de Markov Escondidos," Dissertao de mestrado, Universidade Federal de Campina Grande, Julho 2012.

[6] S. S. Park and N. S. Kim, "On using multiple models of automatic speech segmatation," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2202–2212, November 2007.

Table 1. Distribution of the errors for speech segmentation.

Locution	Total of Boundaries	$\leq$ 10 ms	$\leq$ 20 ms	Segmentation Rate (%)	False Boundaries	No Detection
L01	48	29 (76.3%)	9 (23.7%)	79.16	9	10
L02	22	11 (78.6%)	3 (21.4%)	63.63	6	8
L03	23	11 (68.7%)	5 (31.3%)	69.56	9	7
L04	18	12 (80.0%)	3 (20.0%)	83.33	7	3
L05	41	19 (73.0%)	7 (27.0%)	63.41	5	15
L06	41	24 (70.0%)	10 (30.0%)	82.92	15	7
L07	27	14 (77.8%)	4 (22.2%)	66.66	8	9
L08	34	24 (92.3%)	2 (7.7%)	76.47	4	8
L09	43	28 (82.3%)	6 (17.7%)	79.06	4	9
L10	20	14 (73.7%)	5 (26.3%)	95.00	6	1
L11	21	16 (88.9%)	2 (11.1%)	85.71	5	3
L12	21	10 (66.7%)	5 (33.3%)	71.42	6	6
L13	35	24 (88.9%)	3 (11.1%)	77.14	6	8
L14	29	17 (77.3%)	5 (22.7%)	75.86	8	7
L15	31	17 (77.3%)	5 (22.2%)	70.96	6	9
L16	33	18 (66.7%)	9 (33.3%)	81.81	9	6
L17	37	24 (82.7%)	5 (17.3%)	78.37	4	8
L18	31	18 (85.7%)	3 (14.3%)	67.74	6	10
L19	34	16 (69.5%)	7 (30.5%)	67.64	10	11
L20	25	10 (62.5%)	6 (37.5%)	64.00	5	9
L21	27	18 (94.7%)	1 (5.3%)	70.37	8	8
L22	31	10 (62.5%)	6 (37.5%)	51.61	10	15
L23	28	12 (60.0%)	8 (40.0%)	71.42	5	8
L24	25	13 (65.0%)	7 (35.0%)	80.00	7	5
L25	27	13 (81.2%)	3 (18.8%)	59.25	9	11
L26	30	16 (64.0%)	9 (36.0%)	83.33	6	5
L27	24	12 (63.1%)	7 (36.9%)	79.16	7	5
L28	15	9 (81.8%)	2 (18.2%)	73.33	4	4
L29	17	7 (77.8%)	2 (22.2%)	52.94	4	8
L30	33	17 (73.9%)	6 (26.1%)	69.69	5	10
L31	12	7 (87.5%)	1 (12.5%)	66.66	4	4
L32	25	17 (89.5%)	2 (10.5%)	76.00	9	6
L33	26	13 (72.2%)	5 (27.8%)	69.23	9	8
L34	37	25 (83.3%)	5 (16.7%)	81.08	7	7
L35	30	19 (86.4%)	3 (13.6%)	73.33	8	8
L36	31	22 (84.6%)	4 (15.4%)	83.87	3	5
L37	34	14 (66.7%)	7 (33.3%)	61.76	14	13
L38	23	12 (80.0%)	3 (20.0%)	65.21	7	8
L39	26	17 (85.0%)	3 (15.0%)	76.92	6	6
L40	28	15 (71.4%)	6 (28.6%)	75.00	3	7
L41	30	14 (73.6%)	5 (26.4%)	63.33	5	11
L42	30	17 (80.9%)	4 (19.1%)	70.00	11	9
L43	22	8 (88.9%)	1 (11.1%)	40.90	6	13
L44	14	8 (100.0%)	0 (0%)	57.14	3	6
L45	23	11 (73.3%)	4 (26.7%)	65.21	8	8
L46	28	15 (68.2%)	7 (31.8%)	78.57	6	6
L47	27	22 (91.6%)	2 (8.4%)	88.88	6	3
L48	26	17 (80.9%)	4 (19.1%)	80.76	6	5
L49	36	21 (80.7%)	5 (19.3%)	72.22	7	10
L50	24	15 (75.0%)	5 (25.0%)	83.33	6	4
Average	28.06	77.6%	22.4%	72.41	6.74	7.6

[7] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1444–1449, July 2011.

[8] C. Lin and J. R. Jang, "Automatic Phonetic Segmentation by Score Predictive Model for the Corpora of Mandarin Singing Voices," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2151–2159, September 2007.

[9] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort, "Automatic phone alignment: A comparison between speaker-independent models and models trained on the corpus to align," *Springer Berlin Heidelberg*, vol. 7614, pp. 300–311, 2012.

[10] Eren Akdemir and Tolga iloglu, "Using Visual Information in Automatic Speech Segmentation," *Signal Processing, Communications and Applications Conference*, 2008.

[11] B. Sudhakar and R. Bens Raj, "Automatic Speech Segmentation to Improve Speech Synthesis Performance," *International Conference on Circuits, Power and Computer Technologies (ICCPCT 2013)*, pp. 835–839, 2013.

[12] D. C. Costa and G. A. M. Lopes and C. A. B. Mello and H. O. Viana, "Speech and Phonema Segmantation Under Noisy Environment Through Spctrogram Image Analysis," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1017–1022, October 2012.

[13] B. Zilko and S. Manandhar and R. C. Wilson and M. Zilko, "Wavelet Method of Speech Segmentation," *Proceedings of 14th European Signal Processing Conference EUSIPCO*, 2006.