Study on Unit-Selection and Statistical Parametric Speech Synthesis Techniques

Mohammad Savargiv[a*], Azam Bastanfard[b]

[a] Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[b]Faculty of Media Engineering, Islamic Republic of Iran Broadcast University, Tehran, Iran

Abstract

One of the interesting topics on multimedia domain is concerned with empowering computer in order to speech production. Speech synthesis is granting human abilities to the computer for speech production. Data-based approach and process-based approach are the two main approaches on speech synthesis. Each approach has its varied challenges. Unit-selection speech synthesis and statistical parametric speech synthesis are two dominant speech synthesizer techniques. The naturalness is the main challenge of all speech synthesis approaches. The Intonation, speech style and emotional state are included in naturalness factor and all of them are considered as suprasegmental features. Equipped synthesized speech with paralinguistic information is more believable from the perceptual aspect. Prosody information plays an important role on the synthesized speech quality of text to speech systems. The first purpose of modern speech synthesizer systems is text to speech conversion and the second purpose is transferring the emotional states of text in the voice form. In this paper two main speech synthesis approaches and their challenges are investigated in detail.

Keywords: Speech Synthesis Method, Prosody, Review.

1. Introduction

The input of speech synthesizer systems is text or phonetics symbols and the output is corresponding voice. There are diverse applications of speech synthesis techniques such as text to speech synthesis for foreign language teaching applications, pronunciation teaching applications, audio dictionary, book reader for the blind, robotics, human-computer interaction, etc.

The main goals of speech synthesizer systems are intelligibility and naturalness output. Each one of these goals is multi-dimensional factor. Intelligibility factor refer to the understanding speech by human or automatic speech recognition (ASR) algorithms. The purpose of naturalness is the similarity between human speech and synthesizer output. The naturalness factor is a qualitative factor and it has own indicators such as continues speech, intonation and prosody.

Speech production by synthesizer and transmit both emotional state and personality mode are highlights of speech synthesis [1]. Furnishing synthesized speech with the suprasegmental features lead to naturalness speech production [2]. There are diverse ideas for speech synthesis. Major of them are unit-selection speech synthesis and statistical parametric speech synthesis [3].

The rest of paper is organized as follow. Speech synthesis methods are investigated in the section 2.

* Corresponding author. Email: savargiv@qiau.ac.ir

Section 3 reviewed the challenges of speech synthesis methods and the conclusion and future direction are located in section 4.

2. Review on Speech Synthesis Methods

In this section two major speech synthesis methods are surveyed and diverse prosody models in which improve the naturalness factor are studied.

2.1 Unit-Selection Synthesis (Concatenative Synthesis)

In the unit-selection synthesis pre-recorded units of speech are selected form a repository and placed one after another according to target sentence and play with appropriate rate finally. The repository has large size usually. The simple manner of method caused it become dominant speech synthesis method in the artificial speech production domain. The quality of synthesized speech is directly influenced by the quality of stored speech units. This method has high economic justification and it is useful method for synthesizer applications in which required restricted words. Unit-selection speech synthesis has no reasonable justification for general applications, because it is required to high volume and high diverse speech units in which include heavy cost of production and maintenance. Fig. 1 shows general scheme of unit-selection synthesis.

2.2 Statistical Parametric Speech Synthesis

Statistical parametric speech synthesis is another speech synthesis method with high flexibility. This ability is borrowed from the statistical process models. This model has diverse advantage and disadvantage investigated as follow [3].

a) Ability to Generate Speaker's Audio Features

The highest advantage of statistical parametric speech synthesis is ability and flexibility to change the characteristics of the speaker's voice (e.g. speech style and

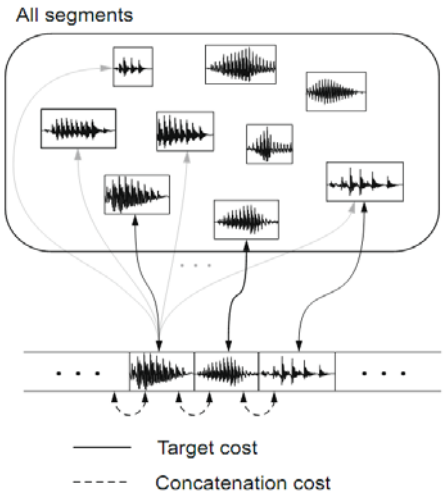

Fig. 1. General Scheme of Unit-Selection Synthesis [3]

speech emotional state).In the statistical parametric speech synthesis, due to change in the parameter models, it is easy to change the characteristics of speaker's voice.

In this regard, there are four main techniques named as Adaptation, Interpolation, Eigen-Voice and Multiple Regression. Adaptive technique is based on voice imitation. It is used to imitation of vocal models of a particular speaker to increasing recognition accuracy. Maximum a Posteriori Estimation (MAP) and Maximum Likelihood Linear Regression (MLLR) are two approaches of the adaptive technique. Voice combination is the other name of interpolation technique. It is able to produce voice and its features at the same time. The voice conversion technique uses interpolation to convert voice. Two mentioned techniques are required to high volume of training data and they are including complicated calculation. Using eigen-voice and eigen-value which reduce training data is useful option. However for any of eigen-voices there is no physical correspondence in the voice. Poor ability to control audio features such as expression style and emotional state is disadvantages of the previous techniques.

There is L-dimensional vector in the multiple regression technique for voice control. L consists of the voice characteristics. Each element of the z series captures the feature of the L series. Combination of four mentioned techniques provides diverse style speech without large speech dataset.

$$Z= [z_1, z_2, \dots , z_L] \quad (1)$$

b) Phonetic Space Coverage

The second advantage of statistical parametric speech synthesis is phonetic space coverage. The main idea of this method is combining units to producing diverse and more speech units. In this case producing naturalness and continuous speech is possible. Although the statistical parametric speech synthesis produces diverse speech data, nevertheless new speech unit production is restricted to speech data of database.

c) Multilingual Support

Supporting multilingual is feasible in the statistical parametric speech synthesis. For this task it is need to determine factors content in any languages. Statistical parametric speech synthesis can initiate by the few training data. The first step of creating speech synthesis application in all languages is collecting speech data. Speech data include hours of voice. The optimal solution is using multilingual idea.

Table 1

Advantages and Disadvantages of Speech Synthesis Methods

Method	Advantages	Disadvantages
Unit-Selection	Simplicity Suitable for Applications with Limited Range Unlimited Duration of Speech Units	The Need to Large Volume Speech Data The Need to Determine the Size of Speech Units The Need to Trade off Between Speech Quality and Speech Data Volume Fragmented Speech Production The Need to Identifying and Selecting the Appropriate Speech Unit
Statistical Parametric	Controllability Flexibility to Producing Various Speech No Need to Use High Volume of Training Data Phonetic Space Coverage The Ability to Produce Continuous Speech Multilingual Support Ability to Use Speech Recognition Techniques	Low Quality Lack of Physical Correspondence for Eigen-Voice

Table 2

Comparison of Speech Synthesis Methods

	Data Volume	Quality	Speech Style Coverage	Control	General Applications Justification
Unit-Selection	High	Influenced by Speech Data	Influenced by Speech Data	Directly Related to the Amount of Data	No
Statistical Parametric	Low	Buzzy Output	Strong	Possibility of Strong Control	Yes

In the statistical parametric speech synthesis instead of using pattern of speech unit, the statistical parameters of speech unit are used. It is the main difference between unit-selection synthesis based on clustering and statistical parametric speech synthesis. In the statistical parametric speech synthesis (e.g. HMM) distribution of fundamental frequency (F0) and duration are clustered independently [5]. Therefore for each features, there are separate decision tree [8]. While in the unit-selection speech synthesis each leaves of tree should be contain waveform of speech. Table 1 and 2 illustrate the most prominent advantages and disadvantages of mentioned speech synthesis methods.

2.3 Expressive Speech Synthesis

There are three categories of effective parameters in the voice. The first category consists of prosodic parameter. The prosodic information plays an important role in the speech synthesizer [2]. The main components of prosody are duration, intonation, phrasing and fundamental frequency. The Rule-based approach and the corpus-based approach are two major prosodic modeling approaches. Linguistic experts extract prosody rule form the natural speech in the rule-based approach. While in the corpus-based approach each speech corpus designed individually and investigates by prosody information on different levels [4]. The rule-based approach is grammatical model and it is based on implicit or explicit knowledge [5]. Table 3 shows strengths and weaknesses points of these models.

The second category consists of excitation parameters. Excitation parameters refer to the excitation features of voice source. The jitter and the shimmer are samples of excitation parameters which obtained from the segmentation levels. Jitter is the mean of fundamental frequency change from one cycle to another. Jitter show different behaviour at different emotional state. For example happiness state includes high jitter and sadness state contains low jitter. Shimmer is the measure of larynx excitation. It is calculated by change of intensity of the pulses from one cycle to another cycle.

The third category consists of vocal tract parameters. The vocal tract contain of the first frequency to the fifth frequency of voice. These frequencies are named as formants of voice.

There are two approaches for expressive speech synthesis. In the first approach the speech is synthesized in the neutral form and then the prosody added to the neutral speech by using transmission techniques [10-12].

Table 3
Comparison of Prosodic Models

Method	Strengths	Weaknesses
Rule Based Approach	Requires fewer resources	Dependence on natural speech Lack of efficiency on the high-volume of data Inability to use multilingual systems
Statistical Approach	The ability to use large volume of data	Data distribution Inefficiencies in natural data collection The lack of optimal state
Integrative Approach	Combining the benefits of two previous approaches.	Required to other features in the languages which have no diacritics
Content-Based Models	The ability to integrating and modeling prosodic forms The ability to control each prosodic forms separated from others	Low efficiency Inappropriate for speech laboratory
Modeling and Labeling Approach	The availability of rich prosodic information	Inappropriate for speech corpus
Prosodic Approaches Based on HMM	More and better prosodic information Best alternative for selected units The availability of intelligible speech	Abnormal speech

The second approach includes three methods to expressive speech synthesis as follow [1].

a) Expressive Speech Synthesis by Explicit Control

The formant-based speech synthesis and the diphon-based speech synthesis are two main expressive speech syntheses by explicit control. The parameters of the formant-based speech synthesis are tuned manually. In this method the emotional states added to the synthesized speech by fundamental frequency and duration factors.

b) Expressive Speech Synthesis by Playback Approach

In the playback approach expressive speech is synthesized based on emotional speech dataset. Unit-selection speech synthesis and HMM-based methods are the examples of this approach. Expressive speech synthesizer in which they are based on unit-selection technique is required to large dataset.

c) *Expressive Speech Synthesis by Implicit Control*

The implicit control uses interpolation among trained statistical models. Flexibility on different emotional states production and speech style production are advantages of implicit control approach.

3. Challenges of Speech Synthesis

There are various challenges in the speech synthesis domain. The first category consists of the challenges of language. Text to speech conversion method consists of two main modules. The first module is front-end. Text analysis is assigned to the front-end part. The word boundary is determined in this part. The back-end module carries out generated waveform of speech. These two modules are developed separately. Text normalization and text-to-phone conversion are done by front-end and back-end modules respectively.

a) Text Normalization Challenge

Normalizing text refer to the words by identical spelling and different pronunciation. In this case different meanings are perceivable. This is the main text normalization challenge of languages which have no diacritics. How to reading numbers and expanding abbreviations are another text normalization challenges. The ordinal number and the cardinal number which are written in the roman form are related samples.

b) Text-to-Phone Challenge

Using dictionary in which contains all pronunciation of words is the simplest approach of text-to-phone conversion. Although this approach is fast nevertheless the main disadvantage is inability to pronounce a word that is not in the dictionary. On the other hand high volume of memory is required.

The second category consists of challenges related to the synthesizer technique. A sequence of audio corresponding to the text cannot induce continuous speech to the listener, consequently the naturalness factor doesn't meet [4].

3.1 Unit-Selection Synthesis Challenges

There are two main challenges in the unit-selection synthesis which they are theoretically similar. The first challenge is target cost. It is associated with mechanism of selecting the units from the speech dataset. The second challenge is concatenation cost in which refers to combination style of selected speech units. Unit selection cost u_i and required unit t_i are modeled according to equation 2 which j is index of feature. Similarly, the concatenation cost is modeled according to equation 3 which k is the audio and spectral features of speech.

$$c^{(t)}(t_i,u_i)=\sum_{j=1}^{p} w_j^{(t)} c_j^{(t)}(t_i,u_i) \qquad (2)$$

$$c^{(c)}(u_{i-1},u_i)=\sum_{j=1}^{q} w_k^{(c)} c_k^{(c)}(u_{i-1},u_i) \qquad (3)$$

Using clustering method is another issue in which reduces unit selection cost. This process puts similar speech units in the one cluster; consequently all units are available at the necessary situation. Fig. 2 shows general scheme of clustering in the unit-selection speech synthesis. The size of selected speech units is another important issue of the unit-selection speech synthesis. Using long duration speech unit is required to handling high volume speech data. On the other side expression style has high diversity; therefore the volume of speech data is very high. The management of such data volume will not easy.

Fewer connection points in the unit-selection speech synthesis lead to continuous synthesized speech. In this case unit-selection synthesizer has strong naturalness factor. So far previous research activity proposed diverse duration for speech units. They are frame duration, duration based on HMM states, diphon duration and half- diphon duration. Optimum duration of speech unit is application based. The volume of dataset and the boundaries of synthesized speech should be considered in the commercial applications.

Producing naturalness speech facilitated if diverse speech style is available. However, it should be a reasonable trade off between the quality of production and the volume of data that must be stored [6]. Sometimes the volume of speech data is more than dozens of hours nevertheless the quality and naturalness is not as expected.

3.2 Statistical Parametric Speech Synthesis Challenges

The quality of synthesized speech is the main challenge of statistical parametric speech synthesis. In general, three

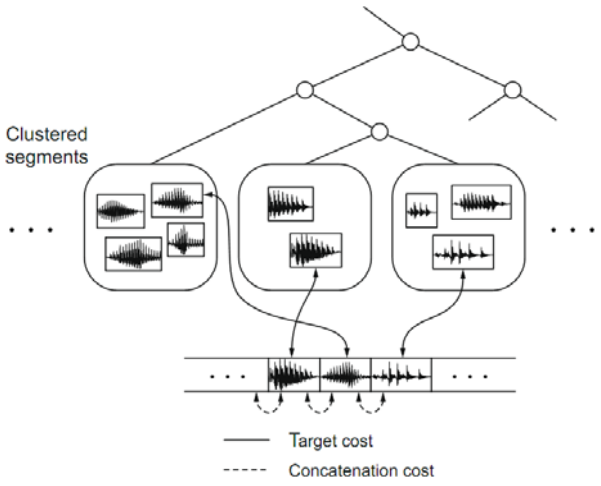

Fig. 2. General Scheme of Clustering in the Unit-Selection Speech Synthesis [3]

main factors are involved with the speech quality.

a) Vocoder

Synthesized speech in the statistical parametric speech synthesis is buzzy because white noise is used in the excitation step.

b) Accuracy of Acoustic Modeling

Speech parameters are produced from the acoustic models. Hence acoustic models are influenced by the quality of speech data. Using dynamic models in which capture speech parameter trajectory is a solution to increasing accuracy of acoustic models.

c) Over Smoothing

Statistical parametric speech synthesis uses speech parameter generation algorithms for generating spectrum parameters and excitation parameters. Using dynamic features in the speech synthesis process provide smooth trajectory for production speech.

4. Conclusion and Future Direction

In this paper speech synthesis techniques and its challenges were investigated from two approaches of unit-selection speech synthesis and statistical parametric speech synthesis. Different techniques of expressive speech synthesis were investigated too.

The presented topics show that statistical parametric speech synthesis provides methods in which they can improve synthesized speech. Speech synthesizer is required to less stored speech data in diverse prosody if using robustness and strong models. Unit-selection speech synthesis methods are required to high volume of speech data in various prosody, speech style and emotional state. While statistical parametric speech syntheses techniques are able to generate models in compound and adaptive form. In the statistical parametric models the need to instances of each blending textual content mode is resolved.

Although there are various methods of speech synthesis, however there is a gap between the naturalness factor and synthesized speech. This challenge is obvious especially in the tonal languages. As the future work we intend to focus on combination of different speech synthesis methods and prosody models to provide language-independent method for expressive speech synthesis.

References

[1] D. Govind and S.R. Mahadeva Prasanna, “Expressive speech synthesis: a review”, Int J Speech Technol 16, pp. 237–260, (2013).

[2] Y. Xu and S. Prom-on, “Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning”, Speech Communication 57, pp. 181–208, (2014).

[3] H. Zen, K. Tokuda and A. W. Black, “Statistical parametric speech synthesis”, Speech Communication 51, pp. 1039–1064, (2009).

[4] K. C. Rajeswari and P. Maheswari, “Prosody Modeling Techniques for Text-to-Speech Synthesis Systems – A Survey”, International Journal of Computer Applications, Vol. 39, No.16, (2012).

[5] H. Tang, X. Zhou, M. Odisio, M. Hasegawa-Johnson, and T.S. Huang, "Two-stage prosody prediction for emotional text-to-speech synthesis", INTERSPEECH, pp. 2138-2141, (2008).

[6] A. Iida, n. Campbell, F. Higuchi and M. Yasumura, “A corpus-based speech synthesis system with emotion”, Speech Communication 40, pp. 161–187, (2003).

[7] C. Valentini-Botinhao, J. Yamagishi, S. King and R. Maia. “Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion”, Computer Speech and Language 28, pp. 665–686, (2014).

[8] X. J. Xia, Z. H. Ling, Y. Jiang and L. R. Dai, “HMM-based unit selection speech synthesis using log likelihood ratios derived from perceptual data”, Speech Communication 63–64, pp. 27–37, (2014).

[9] S. K. Thakur and k. j. satao, “Study of Various Kinds of Speech Synthesizer Technologies and Expression for Expressive Text to Speech Conversion System”, international journal of advanced engineering sciences and technologies, vol. 8, issue no. 2, pp. 301–305, (2011).

[10] D. Govind, S. R. MahadevaPrasanna, B. Yegnanarayana, “Neutral to Target Emotion Conversion Using Source and Suprasegmental Information”, INTERSPEECH, pp. 2969-2972. ISCA, (2011).

[11] J. Tao, Y. Kang and A. Li, “Prosody Conversion from Neutral Speech to Emotional Speech”, IEEE Transaction on Audio, speech and Language Processing, 14, pp. 1145-1154, (2006).

[12] N. Campbell, W. Hamaza, h. Hog and J. Tao, “Editorial Special Section on Expressive Speech Synthesis”, IEEE Transaction on Audio, Speech and Language Processing, 14, pp. 1097-1098, (2006).

[13] N. Campbell, “Conversational Synthesis and the Need for Some Laughter”, IEEE transaction on Audio, Speech and language Processing, 17(4), pp. 1171-1179, (2006).