# SPEECH SYNTHESIS TECHNIQUES. A SURVEY

*Youcef TABET* and Mohamed BOUGHAZI***

* University of M'hamed Bouguerra Boumerdes, Algeria
y-tabet@umbb.dz
**University Badji Mokhtar Annaba, Algeria
Boughazi_m@yahoo.com

**ABSTRACT**

*The goal of this paper is to provide a short but a comprehensive overview of Text-To-Speech synthesis by highlighting its digital signal processing component. First two rule-based synthesis techniques (formant synthesis and articulatory synthesis) are explained then the concatenative synthesis is explored. Concatenative synthesis is simpler than rule-based synthesis, since there is no need to determine speech production rules. However, it introduces the challenges of prosodic modification to speech units and resolving discontinuities at unit boundaries. Prosodic modification results in artifacts in the speech that make the speech sound unnatural. Unit selection synthesis, which is a kind of concatenative synthesis, solves this problem by storing numerous instances for each unit with varying prosodies. The unit that best matches the target prosody is selected and concatenated. To resolve mismatches speech synthesis system combines the unit-selection method with Harmonic plus Noise Model (HNM). This model represents speech signal as a sum of a harmonic and noise part. The decomposition of speech signal into these two parts enables more natural sounding modifications of the signal. Finally Hidden Markov model(HMM) synthesis combined with an HNM model is introduced in order to obtain a Text-To-Speech system that requires smaller development time and cost.*

## 1. INTRODUCTION

In our society, where speed and efficiency are key qualities, a human computer interaction via speech is of great pertinence. Such an interaction involves speech recognition and speech synthesis. The first one consists in extracting the message information in a speech signal so as to control the actions of a machine in response to spoken commands, whereas the second one is the process of creating a synthetic replica of a speech signal so as to transmit a message from a machine to a person, with the purpose of conveying the information in the message [1].

In speech synthesis, the aim is to obtain a synthesized speech not only easily understandable, but also indistinguishable from that produced by a human, in other words, to create a system that equals the human performance. Thus, the two qualities required by a speech synthesis system are intelligibility and naturalness.

There are three main approaches to speech synthesis: formant synthesis, articulatory synthesis, and concatenative synthesis. Formant synthesis models the frequencies of speech signal. Formants are the resonance frequencies of the vocal tract. The speech is synthesized using these estimated frequencies. Articulatory synthesis generates speech by direct modeling of human articulator behavior. On the other hand, concatenative speech synthesis produces speech by concatenating small, prerecorded units of speech, such as phonemes, diphones and triphones to construct the utterance. In case that not just one, but hundreds of realizations of each phonetic speech unit are present in an inventory, a unit selection process must take place in order to create the final synthetic unit sequence. Such speech synthesis method is also called corpus based speech synthesis. The combination of the unit selection with the HNM synthesis gives better quality results. The HMM (Hidden Markov model) combined with HNM (Harmonic plus Noise Model) synthesis system model is introduced in order to obtain less memory to store the parameters of the models and more variations are allowable.

The remainder of the paper is organized as follows. First we give an overview of the Formant synthesis. After that the Articulatory synthesis is depicted. In the fourth part, the Concatenative synthesis is described. This is followed by a survey of the Unit Selection synthesis. The sixth part, describes the HNM (Harmonic plus Noise Model) synthesis. Finally, the HMM (Hidden Markov model) synthesis is shown. The conclusion summarizes the gained experience with using the proposed synthesis techniques.

978-1-4577-0690-5/11/$26.00 ©2011 IEEE

## 2. FORMANT SYNTHESIS

In formant synthesis, the basic assumption is that the vocal tract transfer function can be satisfactorily modeled by simulating formant frequencies and formant amplitudes. The synthesis thus consists of the artificial reconstruction of the formant characteristics to be produced. This is done by exciting a set of resonators by a voicing source or noise generator to achieve the desired speech spectrum, and by controlling the excitation source to simulate either voicing or voicelessness. The addition of a set of anti-resonators furthermore allows the simulation of nasal tract effects, fricatives and plosives. The specification of about 20 or more such parameters can lead to a satisfactory restitution of the speech signal. The advantage of this technique is that its parameters are highly correlated with the production and propagation of sound in the oral tract. The main current drawback of this approach is that automatic techniques of specifying formant parameters are still largely unsatisfactory, and that consequently, the majority of parameters must still be manually optimized [2].

The formant synthesis doesn't use any human speech samples but relies on rules written by linguists to generate the parameters that will permit the synthesis of speech, and to deal with the transition from one phoneme to another, that is, the coarticulation. To write the rules, linguists have studied spectrograms and derived the rules of evolution of formants. However we do not yet know the optimal rule to do this [3]. Moreover, the speech waveform is naturally produced in such a complex process that, currently, rules can only model the features of the speech waveform.

Therefore, the synthesized speech has an artificial, robotic sound, and the goal of naturalness is not reached. However, the rule-based synthesized speech is very intelligible, even at high speeds, which is quite useful for visually impaired for quickly navigating computers using a screen reader. Moreover, when memory and processing costs are limited, such as in embedded systems, these synthesizers are more interesting because they don't have a database of speech samples.

The formant synthesis approach has been implemented in MITalk [4, 5], in KlatTalk [6], and in DECTalk [7].

## 3. ARTICULATORY SYNTHESIS

Articulatory synthesis generates speech by direct modeling of the human articulator behavior, so in principle it is the most satisfying method to produce high-quality speech. In practice, it is one of the most difficult methods to implement. The articulatory control parameters include lip aperture, lip protrusion, tongue tip position, tongue tip height, tongue position and tongue height [8]. There are two difficulties in articulatory synthesis. The first difficulty is acquiring data for articulatory model. This data is usually derived from X-ray photography. X-ray data do not characterize the masses or degrees of freedom of the articulators [3]. The second difficulty is to find a balance between a highly accurate model and a model that is easy to design and control. In general, the results of articulatory synthesis are not as good as the results of formant synthesis or the results of concatenative synthesis.

## 4. CONCATENATIVE SYNTHESIS

The main limitation of formant synthesis and articulatory synthesis is not so much in generating speech from parametric representation, but the difficulty is in finding these parameters from the input specification that was created by the text analysis process. To overcome this limitation, concatenative synthesis follows a data driven approach. Concatenative synthesis generates speech by connecting natural, prerecorded speech units. These units can be words, syllables, half-syllables, phonemes, diphones or triphones. The unit length affects the quality of the synthesized speech. With longer units, the naturalness increases, less concatenation points are needed, but more memory is needed and the number of units stored in the database becomes very numerous. With shorter units, less memory is needed, but the sample collecting and labeling techniques become more complex [9].

The most widely used units in concatenative synthesis are diphones. A diphone is a unit that starts at the middle of one phone and extends to the middle of the following one. Diphones have the advantage of modeling coarticulation by including the transition to the next phone inside the diphone itself. The full list of diphones is called diphone inventory, and once determined, they need to be found in real speech. To build the diphone inventory, natural speech must be recorded such that all phonemes within all possible contexts (allophones) are included, then diphones must be labeled and segmented. Once the diphone inventory is built, the pitch and duration of each diphone need to be modified to match the prosodic part of the specification.

## 5. UNIT SELECTION SYNTHESIS

In concatenative synthesis, diphones must be modified by signal processing methods to produce the desired prosody. This modification results in artifacts in the speech that can make the speech sound unnatural. Unit selection synthesis (also, called corpus-based concatenative synthesis) solves this problem by storing in the unit inventory multiple instances of each unit with varying prosodies. The unit that matches closest to the target prosody is selected and concatenated so that prosodic modifications needed on the selected unit is either minimized or not necessary at all. Since multiple instances of each unit are stored in the unit inventory, a unit selection algorithm is needed to choose the units that best match the target specification. This selection is based on minimizing two types of cost functions, which are target cost and join cost.

In the case of automatic unit selection, the coarticulatory influence isn't limited to the last phoneme.

The database is much larger (1-10 hours) and comprises several occurrences of each acoustic unit, captured under various contexts (like its neighboring phonemes of course, but also its pitch, its duration, its position in the syllable, etc.). As a result, the sequence of phonemes to synthesize leads to a lattice of acoustic units, in which the best corresponds to the expected contexts (prosody, phonetics, etc) but also minimizes the spectral and prosodic discontinuities. Consequently, automatic unit selection requires much less modification of the speech units, which leads to an overall quality of the synthesized speech much more natural than with diphones based synthesis.

Apart from this naturalness, unit selection techniques have several disadvantages. They rely on a very large database, which implies, on the one hand, considerable development time and cost to collect and label the data, and on the other hand, large memory resource requirements to store the data. The second drawback is incorrect labeling and occurrence of unseen target contexts lead to fragments of synthesized speech of extremely poor quality. This phenomenon of unseen contexts may well never be fully overcome with concatenative synthesis as [10] suggest that rare events will always occur in language.

## 6. HNM SYNTHESIS

Prosodic modifications of speech are needed for high quality speech synthesis. HNM models are parametric models, and so it is easy to modify prosodic features like the intonation, stress or rhythm within them with good quality.

A HNM model was first presented in [11]. HNM assumes that the speech signal is composed of a harmonic and a noise part. The harmonic part responds to the quasi-periodic components of the speech and the noise part responds to non-periodic components. These two components are separated in the frequency domain by a time-varying parameter called maximum voiced frequency Fm. The bandwidth up to Fm is represented by harmonic sinusoids and the bandwidth from Fm is represented by a modulated noise component. Unvoiced parts of speech are represented only by noise part. The speech signal is obtained as a sum of the harmonic and the noise part.

The harmonic part contains only harmonic multiples of fundamental frequency. The noise part can be modeled by coding spectral envelope using AR filter, where the synthesis is done by filtering white noise by the AR filter. Since the noise part has no fundamental frequency, the F0 is set to 100 Hz as stated in [11] . The phases of sinusoids are set randomly because the noise is a stochastic signal.

[12] describes and compares three versions of a Harmonic plus Noise Model, HNM, for speech decomposition. The periodic (or quasi-periodic) part is supposed to be harmonic. For the first version of HNM, the harmonic part designates sums of harmonically related sinusoidal components with constant amplitudes within each analysis frame. The phase is modeled by a first-order polynomial (i.e., is assumed to be linear). For the second version, the periodic part is also a sum of harmonically related sinusoidal components, however, with piece-wise linearly varying complex amplitudes. The third version makes use of a p-th order polynomial with real coefficients for the harmonic amplitudes, and the phase is assumed to be linear. Given the harmonic part, the non-periodic part is obtained by subtracting the harmonic part from the original speech signal. The non periodic part (or residual signal) thus accounts for everything in the signal that is not described by harmonic components. It includes the friction noise, the period-to-period fluctuations produced by the turbulences of the glottal airflow, etc.

## 7. HIDDEN MARKOV MODEL SYNTHESIS

In unit selection synthesis, multiple instances of each phone in different contexts are stored in the database. To build such a database is a time consuming task and the database size increases in an enormous way. Another limitation of the concatenative approach is that it limits us to recreate what we have recorded. An alternative is to use statistical parametric synthesis techniques to infer specification to parametric mapping from data. These techniques have two advantages: firstly, less memory is needed to store the parameters of the models than to store the data itself. Secondly, more variations are allowable for example; the original voice can be converted into another voice.

One of the most usable statistical parametric synthesis techniques is the hidden Markov model (HMM) synthesis. It consists of two main phases, the training phase and the synthesis phase. At the training phase, it should be decided which features the models should be trained for. Mel frequency cepstral coefficients (MFCC) and their first and second derivatives are the most common types of features used. The feature are extracted per frame and put in a feature vector. The Baum-Welch algorithm is used with the feature vectors to produce models for each phone. A model usually consists of three states that represent the beginning, the middle and the end of the phone. The synthesis phase consists of two steps: firstly, the feature vectors for a given phone sequence have to be estimated. Secondly, a filter is implemented to transform those feature vectors into audio signals.

The quality of the HMM generated speech is not as good as the quality of the speech generated from unit selection synthesis. The modeling accuracy can be improved by using hidden semi-Markov models (HSMMs) [13], trajectory HMMs [14], and stochastic Markov graphs [15].

[16] integrates the Harmonic plus Noise Model (HNM) into the Hidden Markov Model-based Speech Synthesis System (HTS). This integration leads to a Text-To-Speech system that requires smaller development time and cost, in comparison with the usual state-of-the-art Text-To-Speech systems typically based on automatic selection and synthesis of sub-words units (e.g., diphones), while also producing a better quality speech

output (compared to HTS alone). This quality enhancement is achieved by replacing the source filter modeling approach typically used in HTS with the HNM model, which is known for being able to synthesize natural sounding speech under various prosodic modifications.

## 8. CONCLUSION

In This paper, we have presented a survey of several speech synthesis techniques. Formant synthesis and articulatory synthesis are less used today but these techniques can be suitable for applications that require less memory and low processing cost. The focus nowadays is on the unit selection synthesis combined with harmonic plus noise model (HNM). Such synthesis methods allow for more natural-sounding modifications of the signal. The parametric representation of speech using HNM provides a straightforward way of smoothing discontinuities of acoustic units around concatenation points. The main limitation of the unit selection synthesis combined with HNM is high processing cost and fewer variations are allowable on the recorded data. Hidden Markov Model Synthesis is statistical methods that allow more variations on th recorded data and this method will become dominant. The integration of the Harmonic plus Noise model into the Hidden Markov model-based speech synthesis system leads to a Text-To-Speech system that requires smaller development time and cost.

## REFERENCES


[1] L.R Rabiner, "Applications of voice processing to telecommunications," Proc. IEE, vol. 82, pp. 199-228, 1994.

[2] T .Styger, & E. Keller. "Formant synthesis," In E. Keller (ed.), Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges (pp. 109-128). Chichester: John Wiley.1994

[3] D.H. Klatt "Review of text-to-speech conversion for English," Journal of the Acoustical Society of America, vol. 82(3), 1987.

[4] J. Allen, S. Hunnicutt, R. Carlson and B. Granstrom, "MITalk-79: The 1979 MIT text-to-speech system," Speech communications papers presented at the 97th meeting of the acoustical society of america, Cambridge, USA, pp. 507-507,1979 .

[5] J. Allen, S. Hunnicutt and D.H. Klatt, "From Text-to-speech: The MITalk System," Cambridge University Press, Cambridge, 1987.

[6] D.H. Klatt, "The klattalk text-to-speech conversion system," Proceeding on the international conference on acoustic, speech and signal processing, Paris, pp. 1589-1592, 1982.

[7] D.H. Klatt, "DecTalk user's manual," Digital Equipment Corporation Report, 1990.

[8] B. Kroger , "Minimal Rules for Articulatory Speech Synthesis," Proceedings of EUSIPCO92, pp.331-334,1992.

[9] T. Dutoit, "High-Quality Text-to-Speech Synthesis: an Overview," Journal of Electrical & Electronics Engineering,Australia: Special Issue on Speech Recognition and Synthesis, vol. 17, pp. 25-37, 1999.

[10] B. Mobius, "Rare events and closed domains: Two delicate concepts in speech synthesis," Proceedings of the 4th ESCA Workshop on Speech Synthesis, Perthshire, Scotland, 2001.

[11] Y. Stylianou, "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification," PhD thesis, Ecole Nationale Supérieure des Telecommunications, Paris,January 1996

[12] Y. Stylianou, "Modeling Speech Based on Harmonic Plus Noise Models," Springer, 2005

[13] K. Tokuda et al., "Hidden semi-Markov model based speech synthesis," in Inter speech, pp. 1185–1180, 2004.

[14] K. Tokuda et al., "An introduction of trajectory model into HMM-based speech synthesis," in ISCA SSW5, 2004.

[15] M. Eichner et al., "Speech synthesis using stochastic Markov graphs," in ICASSP, pp. 829–832, 2001.

[16] C. Hemptinne. "Integration of the Harmonic plus Noise Model into the Hidden Markov Model-Based Speech Synthesis System (HTS)," Master Thesis:IDIAPLauzane, Suisse 2006.