A Hybrid Method Oriented to Concatenative Text-to-Speech Synthesis

Ignasi Iriondo, Francesc Alías, Javier Sanchis, Javier Melenchón

Department of Communications and Signal Theory
Enginyeria i Arquitectura La Salle
Ramon Llull University, Barcelona, Spain
{iriondo, falias, xsanchis, jmelen}@salleURL.edu

Abstract

In this paper we present a speech synthesis method for diphone-based text-to-speech systems. Its main goal is to achieve prosodic modifications that result in more natural-sounding synthetic speech. This improvement is especially useful for emotional speech synthesis, which requires high-quality prosodic modification. We present a hybrid method based on TD-PSOLA and the harmonic plus noise model, which incorporates a novel method to jointly modify pitch and time-scale. Preliminary results show an improvement in the synthetic speech quality when high pitch modification is required.

1. Introduction

A text-to-speech (TTS) system converts an input text into a speech signal simulating the human text read by a machine.

Most TTS systems are composed of two modules [1]: a) The natural language processing (NLP) module that converts the input text into a sequence of phonemes with their associated prosodic information (energy, duration and pitch, typically). This task requires different sub-modules depending on the language. Text pre-processing, accentuation, syllabic decomposition, phonetic transcription, morphologic and syntactic analysis are necessary to generate the phonetic and prosodic information properly. b) The digital signal processing (DSP) module synthesizes speech from the information generated by the NLP module. Basically, synthesis methods can be classified in rule-based synthesis or synthesis by unit concatenation.

The quality of a TTS system can be measured in terms of the intelligibility and the naturalness of the synthetic speech. Nowadays, the intelligibility rate of the majority of systems is very high; however, there are a lot of potential applications that would require better naturalness. A bad modeling and/or modification of the prosody may cause lack of naturalness. The prosody automation is a complex task because it involves information related to the speaker and linguistics. Firstly, prosody depends on the emotional state, the age, the gender or the geographical origin of the speaker. Secondly, it depends on the sentence type (declarative, interrogative, imperative or exclamative) and its syntactic structure. Therefore, the prosodic modification is a critical aspect of the speech synthesis systems.

The main contribution of this work is the development of a new technique to achieve good prosodic modification when synthesizing emotional speech. This paper explains the DSP module of our Catalan and Castilian Spanish TTS systems. This method, based on the Pitch Synchronous Overlap and Add (PSOLA) [2] improves the prosodic modification by means of a harmonic plus noise parameterization [3] and a novel method to modify time-scale.

2. Related Work

In the concatenative speech synthesis framework, TD-PSOLA [2] and similar methods have been the most used techniques to generate synthetic speech. These methods perform a pitch-synchronous analysis and synthesis of speech. Pitch and time-scale modifications [4] are carried out by means of non-parametric techniques that achieve good performance with a very low computational cost. However, it has some drawbacks that have been analyzed in different works such as [5], [6], [7]. These and other authors propose methods to overcome TD-PSOLA problems. In [8], Y. Stylianou presents the application of harmonic plus noise models (HNM) to TTS synthesis. This parameterization is based on the decomposition of the speech into a harmonic component plus a noise component (see equation (1)). Voiced speech is formed by a harmonic part plus a noise part, while unvoiced speech is only formed by the noise component.

$$s(t) = h(t) + n(t) \tag{1}$$

The harmonic part (equation (2)) is a finite sum of harmonic signals of the fundamental frequency, which models the *quasi-periodic* part of the speech. The complex amplitude of every harmonic signal is calculated to match as closely as possible the original signal. These time-varying vectors of complex amplitudes represent the voiced spectrum of the speech. The number of harmonics is also time-varying and it will be another parameter of the model.

$$h(t) = \sum_{k=1}^{K(t)} A_k\,(t)\, e^{j2\pi k f_o t} \tag{2}$$

The noise part models the non-periodic component of the speech such as the friction noise and the unvoiced sounds. This part is obtained from (1) and it can be parameterized following different techniques. For instance, in HNM1 [8], an AR model is used to estimate this spectral band.

Figure 1 shows the spectrum of a vowel divided in two bands that are separated by the maximum voiced frequency (MVF). The spectral band below the MVF contains the harmonic part and the spectrum above the MVF represents the major contribution of the noise component.

3. Our Concatenative TTS System

3.1. The Speech Corpus

The speech corpus recording, labeling and parameterization are critical tasks related to the synthetic speech quality. Our speech corpus is composed of diphones and triphones. We have chosen 37 phones for Catalan that would generate 1369 combinations

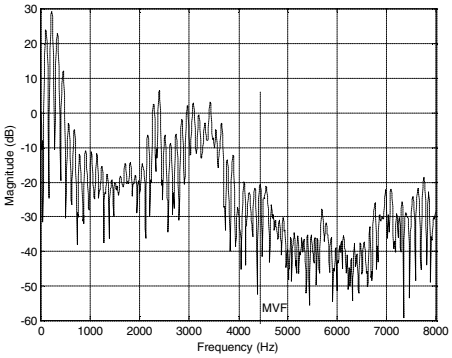

Figure 1: Voiced speech spectrum

of pairs, but not all can be pronounced in Catalan. In order to cover this phonetic content, we have chosen 895 diphones and two types of triphones. The first type is [C + /r/,/l/,/w/ + V] and it is necessary due to the short duration of the liquid consonant and the apparition of a short vowel when the liquid consonant is /r/. The second type is [V + /r/ + C] also including this short vowel after /r/. So, the speech corpus includes 312 triphones, making a total of 1207 units. The number of units of the Spanish Corpus is reduced to 698 units (488 diphones and 210 triphones) due to the fact that Spanish has only five vowels, while Catalan has eight. Each unit was inserted into a carrier sentence to be recorded at near-constant pitch by a professional speaker.

Next, a semi-automatic segmentation process was carried out using Hidden Markov Models. In our approach, the full phones are saved in the units. Also, every phone is labeled with five segmentation marks: the beginning and the end of the phone, and the beginning, the center and the end of the stable part. As we can see below, the stable part is used to modify the duration of the phone and the center of the stable part is used to concatenate units. It is important to note that there are some phones that require an accurate process of segmentation (unvoiced plosives, the phoneme /R/, triphones and silents).

Furthermore, our synthesis system uses phonetic information to choose the optimal concatenation point of the two units needed to synthesize a phoneme. This phonetic information consists of the type of phoneme (vowel, semivowel, consonant or semi-consonant), the place and the manner of articulation, and the sonority (voiced or unvoiced). Sonority is necessary due to different synthesis techniques are applied to voiced and unvoiced speech.

3.1.1. Pitch Synchronous Analysis

The units of the speech corpus have been pitch-synchronously analyzed. Therefore, a process for automatically marking pitch pulse locations is used. The implemented algorithm obtains the pitch marks by means of dynamic programming applied to the energy contour of the signal [9]. These marks are placed along the whole signal without any voiced/unvoiced decision.

3.1.2. Voiced/unvoiced decision

Our system does not need an automatic voiced/unvoiced decision because this information is included in the phonetic information mentioned above.

3.1.3. Maximum Voiced Frequency

We have defined a Maximum Voiced Frequency (MVF) for each semiphone which is calculated from its stable part. Therefore, there are two MVF values for diphones and three for triphones.

The computation of the MVF for every speech frame is based on the Multiband Excitation (MBE) model [10]. The short-time spectrum of the analyzed frame is divided into multiple bands, which are centered in the harmonics of the fundamental frequency (F_0) with a F_0 bandwith. Every band is classified as voiced or unvoiced. The MVF is obtained from the multiband sonority after some post-processing steps. Firstly, the searching process is limited to a range of frequencies: $F_{min} \leq$ MVF $\leq F_{max}$. Secondly, the multiband information is expanded on the frequency domain, taking into account the corresponding bandwidth of the harmonics. Finally, the MVF is assigned to the initial frequency of the first unvoiced band of $B \cdot F_0[Hz]$ bandwidth, excluding spurious bands. After several experiments, we have adjusted $B = 4$, $F_{min} = 2000Hz$ and $F_{max} = 5000Hz$. Figure 2 depicts an example of the MVF computation result.

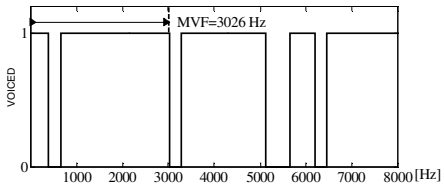

Figure 2: Example of the MVF superimposed on the binary voiced information of a speech frame in the frequency domain.

3.1.4. Harmonic Component Calculation

In the calculation of the amplitudes and phases of the harmonic component, we suppose that both of them are constant for each analysis frame, just like the pitch period and the MVF. We have employed the discrete version of the equation (3), where the amplitude A_k is complex, and therefore, it contains information about the magnitude and the phase of the corresponding harmonic.

$$h(t) = \sum_{k=-L}^{L} A_k \left(t_a^i\right) e^{j2\pi k f_o(t_a^i)(t-t_a^i)} \tag{3}$$

where L is the number of harmonics, A_k is the complex amplitude of the harmonic k and it verifies $A_{-k} = A_k^*$. The calculation of these complex amplitudes is obtained minimizing a weighted time-domain least-squares criterion with respect to A_k:

$$\varepsilon = \sum_{t=t_a^i-N}^{t_a^i+N} w^2(t)\ (s(t) - h(t))^2 \tag{4}$$

where $w(t)$ represents the Hanning window and N is the closest integer to the local pitch period. It is important to emphasize that the analysis window is centered in the analysis time instant t_a^i and the length of the window is $M = 2N + 1$. A_k is estimated by resolving an over-determined system of linear equations [3].

In conclusion, the parameters of the harmonic part are the fundamental frequency, the MVF, and 2L+1 complex amplitudes A_k.

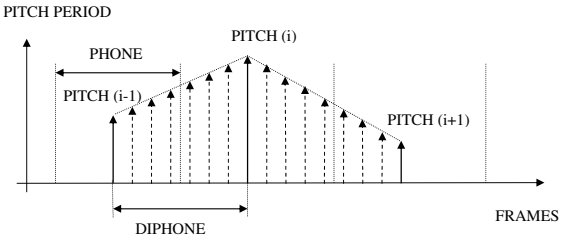

Figure 3: Example of the pitch interpolation between phonemes.

3.1.5. Noise Component

The noise part $n(t)$ of equation (1) is calculated as the difference between the speech signal and the harmonic component for voiced speech frames. For unvoiced speech, it is the original speech waveform. In our first approach, we have not parameterized this component and it is saved in wave format for each unit. Currently, we are working on the parameterisation of this component in order to compress the speech database and improve the speech synthesis.

3.2. Speech Synthesis

This section describes the speech synthesis process, starting from the information generated by the prosodic module. The main goal of this process is to achieve a good quality for the synthetic speech, matching as far as possible the prosodic information.

3.2.1. Unit selection from the database

The presented TTS system is based on diphones and triphones as basic synthesis units. The unit selection system generates the sequence of units to be synthesized from the phoneme transcription of the text. Currently, this module first searches for the largest units (triphones) from left to right, and secondly, it tries to find a shorter unit (diphone). However, this method could also be used in a variable-length unit selection system.

3.2.2. Pitch and duration target

The prosodic module generates a pitch value in Hz and a duration value in milliseconds (ms) for every phoneme to synthesize. These values are closely connected because the adjustment of one of them has an effect on the other. For instance, the number of frames to a given duration will depend on the desired pitch. Therefore, we generate a set of synthesis pitch marks that takes into account the final pitch and duration of every phoneme. The obtained duration is not exactly the desired one (in ms), because the number of frames is an integer and these values will rarely coincide.

Moreover, it is necessary to interpolate the pitch between the current phone and its contextual phones in order to avoid sharp transitions of the pitch. Figure 3 shows the interpolation scheme used to generate the final pitch marks.

Finally, the whole speech wave is generated by means of overlap and add (OLA) of the synthesized frames centered in the synthesis pitch marks. The synthesis process and the frame conversion to adjust the duration are described in detail below.

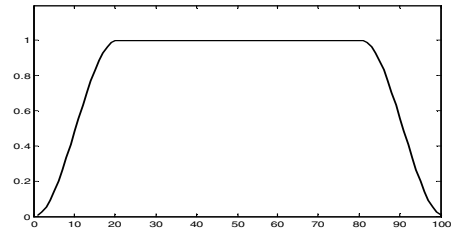

Figure 4: Pseudo-hanning window.

3.2.3. Voiced frame synthesis

Voiced frames have been separated into the harmonic component and the noise component in the analysis process. Now, we need to rebuild these frames, but with a different fundamental frequency, f'_o (equation (5)). Therefore, it is necessary to recalculate the complex amplitudes, $\hat{A}_k$ of the new harmonics without modifying the spectral envelope of the sound. These new amplitudes are estimated by means of an interpolation of the two adjacent original harmonics. The length of the frame is the difference between the adjacent synthesis pitch marks.

$$\hat{h}(t) = \sum_{k=-L}^{L} \hat{A}_k \left(t_s^i\right) e^{j2\pi k f'_o (t_s^i)(t-t_s^i)} \tag{5}$$

Next, we add the noise part weighted by a hanning window that is centered in the synthesis pitch mark. This process is repeated for every frame of both units involved in the phoneme generation.

3.2.4. Unvoiced frame synthesis

Most concatenative speech synthesizers divide unvoiced speech into regular frames (typically 10 ms). The increase of the duration by repetition of frames may cause the sound to become voiced. In order to avoid this undesirable effect, we generate unvoiced phones from only two frames. The left frame joints the transition part with the stable part that has been enlarged or reduced depending on the required duration. The right frame is similarly built. Both frames are weighted by a pseudo-hanning window (see Figure 4) to perform the overlap and add process. The enlargement of the stable part is achieved by means of the repetition of mirrored copies of the original samples. To reduce the duration, samples of the stable part are eliminated [11].

3.2.5. Unit concatenation

Pitch modification is ready once the synthesis instants have been calculated and the frames are synthesized. It is only necessary to adjust the number of frames that matches the final duration. The repetition or elimination of frames to increase or decrease the duration of a phoneme tends to cause a lack of naturalness in the synthesized speech. In order to minimize this effect, the number of frames is converted by means of the method described in [11]. To convert N frames to M frames, with N greater, less or equal to M, a NxM matrix of transformation coefficients is generated. The coefficient a_{ij} determines the ratio of the x_i frame to the y_j frame (see Figure 5). This kind of transformation is only carried out in the stable part, and, therefore, we distinguish between fixed part and variable part. The whole process to match the desired duration of a phoneme is shown in figure 6.

Finally, the synthetic speech signal is generated by means

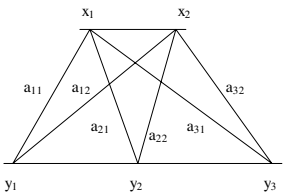

Figure 5: Conversion of N frames into M.

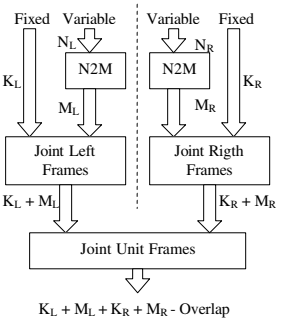

Figure 6: Process to synthesize a phoneme with the required number of phonemes.

of the PSOLA of all the final frames centered in their corresponding synthesis pitch marks.

4. Experiment

We have tested the system with and without the HNM module, to compare the obtained results. First, a copy-synthesis experiment is carried out over a set of nine sentences with their phonetic and prosodic information extracted from a television documental. Informal tests show no significant differences among results due to the input prosodic target does not include considerable variations of pitch.

In a second experiment, the input is a set of three sentences with prosodic information corresponding to four basic emotions (joy, fear, anger, and sadness). In [12], these 12 utterances were performed by means of our PSOLA-based TTS system. These have now been tested with the system proposed in this paper and there is a noticeable improvement when high pitch variability is required (i.e anger).

5. Conclusion

In this paper we present a concatenative speech synthesis system oriented to improve pitch and time-scale modification that is required in text-to-speech synthesis. Both speech analysis and synthesis are pitch-synchronous, and moreover, speech frames are parameterized using a harmonic plus noise decomposition. The final frames that constitute a synthesized phoneme are calculated through a novel method which establishes a fixed part and a variable part of the units. The fixed part is formed by the frames from the transition to the adjacent phoneme and the variable part is the stable segment of the phoneme. The time-scale modification is achieved by a linear combinations of all original frames of the stable part joined with the transition part. On the other hand, pitch modification involves two steps. Firstly, the synthesis instants are calculated and, secondly, the new harmonic complex amplitudes are estimated from the original ones. This parameterized version of TD-PSOLA achieves better pitch modification. In addition, the speech corpus processing and the synthesis process have been summarized.

This synthesis method has been tested with our Catalan and Spanish concatenative TTS systems, and preliminary results show an improvement with regard to the PSOLA implementation. The following steps are the smoothing of inter-frame complex amplitudes and the parameterization of the noise part.

6. Acknowledgements

We would like to thank the Generalitat de Catalunya and the D.U.R.S.I. for their support under grant 2000FI-00679.

7. References

[1] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Dordrecht (Germany): Kluwer, 1997.

[2] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, December 1990.

[3] Y. Stylianou, *Harmonic plus Noise Models for Speech combined with Statistical Methods for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Superieure des Telecommunications, Paris, 1996.

[4] E. Moulines and J. Laroche, "Non-Parametric Techniques for Pitch-Scale Modification of Speech," *Speech Communication*, vol. 16, pp. 175–205, 1995.

[5] T. Dutoit and H. Leich, "MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database," *Speech Communication*, vol. 13, pp. 435–440, June 1996.

[6] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus Harmonic plus Noise Model in diphone based speech synthesis," *ICASSP-98*, 1998.

[7] F. Violaro and O. Böeffard, "A Hybrid Model for Text-to-Speech Synthesis ," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 426–434, September 1998.

[8] Y. Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis," *IEEE Transactions on Speech and audio Processing*, vol. 9, pp. 21–29, January 2001.

[9] F. Alías and I. Iriondo, "Asignación automática de marcas de pitch basada en programación dinámica," *Procesamiento del Lenguaje Natural*, pp. 225–231, September 2001.

[10] D. Griffin and J. Lim, "Multiband-excitation vocoder," *IEEE Transactions Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 236–243, February 1988.

[11] R. Guaus, J. Oliver, H. Moure, I. Iriondo, and J. Martí, "Síntesis de voz por concatenación de unidades: Mejoras en la calidad segmental," *Tecniacústica 98. Lisboa*, September 1998.

[12] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, D. Tena, and L. Longhi, "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques," *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 161–166, September 2000.