Efficient Speech Synthesis System using the Deterministic plus Stochastic Model

Daniel Erro, Asunción Moreno

Department of Signal Theory and Communications
Technical University of Catalonia, Barcelona, Spain
{derro,asuncion}@gps.tsc.upc.edu

Abstract

In this paper, a high-quality concatenative synthesis system using the deterministic plus stochastic model of speech is described, in which the prosodic modifications are performed by means of very simple and efficient operations, as we reported in a previous work [11]. In particular, pitch-synchrony is not necessary, and linear interpolations substitute other types of estimation. The method for the concatenation of units has been improved in order to avoid waveform and spectral mismatches.

1. Introduction

The objectives of the TC-STAR project (Technology and Corpora for Speech to Speech Translation, FP6-506738) are extremely ambitious: to make a breakthrough in Speech to Speech Translation (SST) research and reduce significantly the gap between human and machine performance. New algorithms and methods are being developed in order to integrate the linguistic knowledge into the statistical approach to spoken language translation and the statistical modeling of pronunciation of unconstrained conversational speech in automatic speech recognition. New acoustic and prosodic models are also being studied for generating expressive speech synthesis, intra-lingual and cross-lingual voice conversion.

Within the framework of this project, our intent is to research into voice conversion, which consists of modifying the voice of a speaker (source speaker) to be perceived as if another speaker (target speaker) had uttered it. This task is even more complicated when several languages are involved in it. Thus, it is necessary to work with flexible speech models that allow easy prosodic modification and synthesis, but also with those who provide a solid basis for a voice conversion system.

TD-PSOLA and MBR-PSOLA [1] are two well known methods for voice transformation and synthesis, but their use in voice conversion is not appropriate because they assume no model for the speech signal. In addition, there are comparative studies that affirm that the quality and naturalness in the speech fragment concatenation is lower than in other methods [2]. The speech quality provided by other methods like LPC or residual-excited LPC is not as high as desirable.

The sinusoidal model has been used during the last twenty years for analysis, synthesis, coding and compression of speech signals and music. The model became more popular when McAulay and Quatieri proposed a speech analysis/synthesis system based on a sinusoidal representation [3] and described the implementation of time-scale and pitch-scale modifications [4]. The need of a more complex model that could handle the non-harmonic component of sound led to a deterministic plus stochastic decomposition of speech [5, 6]. The main advantage of such model is that it provides good knowledge of the signal from the perceptual point of view, and allows to manipulate many characteristics of the signal by changing its parameters in a very flexible way.

Many synthesis systems based on sinusoidal models or deterministic plus stochastic models can be found in the literature [4, 7, 8, 9]. What makes them different is the way of performing the analysis and the prosodic modifications. In some of them a pitch-synchronous scheme is used, where the signal is divided into frames containing one or two pitch periods, and the prosodic modifications are performed by changing the distance or the number of replicas of the windowed frames, like in the PSOLA method [7, 8]. A very precise separation of the pitch periods is necessary for the analysis. On the other hand, McAulay and Quatieri use a set of special time instants called onset times [4], which represent the glottal closure instants. They also divide the amplitudes and the phases in two terms, the first one related with the spectral shape of the glottal source and the second one caused by a filter modelling the vocal tract. In [9] it is not necessary to calculate the onset times, but cubic polinomials are used for every phase manipulation, and the method requires the separation between the source and the filter components in order to perform prosodic modifications.

In [11] we presented a tool for speech analysis, modification and concatenative synthesis, in which the prosodic modifications were performed by means of simple and efficient calculations, independently of the time instants where the signals had been analyzed. Special emphasis was placed on the phase manipulation. In this work, some improvements have been done at the concatenacion block of the previous system, so an overview of the full system is presented. The paper is organized as follows: in section 2 the analysis and resynthesis process is described; in section 3 it is explained in detail how the prosodic modifications are carried out; in section 4 the new procedure for the concatenation of units is presented; in sections 5 and 6 some aspects are discussed and the main conclusions are enumerated.

2. Modeling of speech

The deterministic plus stochastic model [5] assumes that the speech signal can be represented as a sum of a number of sinusoids with time-varying parameters and a noise-like component:

$$s[n] = \sum_{j=1}^{J} A_j[n] \cdot \cos\left(\theta_j[n]\right) + e[n] \qquad (1)$$

The deterministic component is present only in the voiced fragments of speech, when the vocal folds vibrate with a certain fundamental frequency. The vibration can be modelled as a locally periodic train of pulses and the vocal tract acts as a filter which can be considered linear, so that it can be assumed that the frequencies of the sinusoids are harmonically related.

The stochastic component contains all the non-sinusoidal signal components: frication, breathing noise, etc. It can be characterized by its local spectral power density.

Both the deterministic and the stochastic component have time-varying parameters that can be considered stable within short intervals. Thus, the signal is analyzed locally by frames.

2.1. Analysis

The first task consists of extracting the fundamental frequency of the signal. The information of the pitch is used to decide whether a speech fragment is voiced or not. It is a binary decision, because it implies the presence or absence of deterministic component. The pitch is considered to be zero in the unvoiced fragments. We have used a set of pitch marks obtained during the recording process from the electroglotographic signal to calculate the pitch contour, but any other pitch detection method is also suitable.

The amplitudes and frequencies of the harmonic sinusoids are measured at the so called *measurement points* located in samples $n = n_k$, $k = 1, 2, 3\ldots$ From now on, the k^{th} measurement point will be called simply *point k*. For simplicity, $n_k = k \cdot N$, and N is a constant number of samples corresponding to a time interval of 8 or 10ms. If the fundamental frequency f_0 is greater than zero at point k, the amplitudes and phases of every harmonic below 5KHz are detected [12].

Once the amplitudes and phases are known, the deterministic waveform is interpolated at every time instant and subtracted from the original sound in order to isolate the stochastic component. Let $A_j^{(k)}$ be the amplitude of the j^{th} harmonic at point k. The instantaneous amplitude of each sinusoid is linearly interpolated between k and k+1 [3]:

$$A_j[kN+m] = A_j^{(k)} + \tfrac{m}{N}\left(A_j^{(k+1)} - A_j^{(k)}\right) \tag{2}$$

for m=0 to m=N–1. The phases and frequencies are interpolated together by a 3rd order polynomial which models the instantaneous phase of each sinusoid, whose derivative is the instantaneous frequency.

$$\theta_j[kN+m] = am^3 + bm^2 + cm + d \tag{3}$$

for m=0 to m=N–1. The coefficients $\{a, b, c, d\}$ are chosen to satisfy the initial and final conditions in an optimal manner [3]. The complete deterministic waveform is finally calculated:

$$d[n] = \sum_{j=1}^{J^{(k)}} A_j[n] \cdot \cos\theta_j[n], \quad \forall n \tag{4}$$

The stochastic component is isolated by subtracting the interpolated deterministic waveform $d[n]$ from the original speech signal. The residual is then analyzed by N-length frames centred at the measurement points, where the interpolation has minimum error. The linear predictive coding (LPC) technique is used to model the magnitude spectral shape of the residual.

2.2. Resynthesis

The speech signal can be resynthesized from its measured parameters and the system output is perceptually equivalent to the original. They are almost indistinguishable. The deterministic component is rebuilt by the overlap-add (OLA) technique [10]. A frame of 2N samples is built at each point k by summing together all the detected sinusoids with constant amplitudes, frequencies and phases. A triangular 2N-length window is used to overlap and add the frames in order to obtain the time-varying synthetic deterministic component, which can be described as follows:

$$\begin{aligned} d[kN+m] = \sum_{j=1}^{J^{(k)}} \Big(& A_j^{(k)} \tfrac{N-m}{N} \cos\left(w_j^{(k)} m + \varphi_j^{(k)}\right) + \\ & + A_j^{(k+1)} \tfrac{m}{N} \cos\left(w_j^{(k+1)}(m-N) + \varphi_j^{(k+1)}\right) \Big) \end{aligned} \tag{5}$$

For the generation of the stochastic component, N-length frames of white gaussian noise are shaped in frequency by the previously calculated LPC filters. The final conditions of the k^{th} LPC filter are used as initial conditions for the $(k+1)^{\text{th}}$ filter.

3. Prosodic modifications

3.1. Duration modification

The duration modification can be carried out by increasing or decreasing the distance N between the different measurement points. The modification factor ρ can be constant over the entire signal, but can also be different at each point k. Thus, the amplitude and fundamental frequency variations are adapted to the new time scale. On the other hand, if the phases were kept unmodified the waveform coherence between consecutive points would be lost. Therefore, the change in N needs to be compensated with a phase manipulation in a way that the waveform and frequency of the duration-modified signal are similar to the original. For this purpose, we developed the method in [11].

Assuming that the fundamental frequency varies linearly from point k–1 to k, the phase evolution of the first harmonic between these two points can be approximated by Ψ:

$$\begin{aligned} & \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) = \\ & = 2\pi T_s \int_0^N \left[f_0^{(k-1)} + \tfrac{n}{N}\left(f_0^{(k)} - f_0^{(k-1)}\right)\right] dn = \\ & = \left(f_0^{(k-1)} + f_0^{(k)}\right) \cdot \pi N T_s \end{aligned} \tag{6}$$

T_s is the sampling period. If N is changed by N', an increment has to be added to the phases to compensate the movement of the measurement point. The increment at the first harmonic can be calculated as:

$$\Delta\varphi_1^{(k)} = \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N'\right) - \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) \tag{7}$$

It can be assumed that this term compensates the effect of the modification of N without affecting other types of phase variation, like for example the small changes in the vocal tract phase response. The correction of the phase of the harmonics is performed as follows:

$$\varphi_j'^{(k)} = \varphi_j^{(k)} + j\sum_{q=1}^{k} \Delta\varphi_1^{(q)} \quad j = 1\ldots J^{(k)} \quad \forall k \tag{8}$$

The phase correction at point k must take into account all the previous correction terms already calculated. As the frequencies are harmonically related, the increment at the j^{th} harmonic is exactly j times the increment at the first one.

The modified stochastic component is easily obtained by filtering longer or shorter frames of white noise with the same LPC coefficients that were calculated during the analysis.

3.2. Pitch modification

When the pitch of the signal is modified, the amplitudes of the sinusoids have to be also modified to keep the spectral envelope of the speaker unaltered. The phases of the new harmonics $l \cdot f'_0$ need to be estimated from the original data. Furthermore, the change in the periodicity of the signal implies that the waveform coherence between adjacent measurement points is lost, because the length of the fundamental period changes while the distance N is kept constant.

In the case of the amplitudes $A_l^{(k)}$, a simple linear interpolation between the measured values $A_j^{(k)}$ in dB is enough [7]. The new amplitudes are multiplied by a gain term, because the new frequency spacing between the harmonics increases or decreases the number of sinusoids inside the bandwidth of analysis, while their energy must be kept constant.

$$A_l'^{(k)} = A_l^{(k)} \cdot \sqrt{f_0'^{(k)} / f_0^{(k)}} \tag{9}$$

In pitch-synchronous systems, the phases of the new harmonics $l \cdot f'_0$ can be obtained by means of a linear interpolation of the real and imaginary parts of the complex amplitudes [7]. The same idea can be used in our system [11], but the interpolation has to be done in the same conditions for all the measurement points, in order to guarantee the coherence. That is the reason why in a first step all the phases are moved to the closest crossing-by-zero of the first harmonic.

$$\Phi_{desp} = \varphi_1^{(k)} \tag{10}$$

$$\beta_j^{(k)} = \varphi_j^{(k)} - j\Phi_{desp} \quad j = 1 \ldots J^{(k)} \tag{11}$$

Now the new phases $\beta_l^{(k)}$ are obtained by linear interpolation of the real and imaginary parts of the complex amplitudes given by $A_j^{(k)}$ and $\beta_j^{(k)}$ [7]. After the interpolation process, the phases are moved back to the original location:

$$\varphi_l^{(k)} = \beta_l^{(k)} + l\Phi_{desp} \quad l = 1 \ldots L^{(k)} \tag{12}$$

The phases for the new harmonics obtained from equation (12) are correct, but the relative position of the measurement point within the pitch period has changed. A new phase term has to be added to compensate the modification of the periodicity. The equation (6) is useful again. The correction is applied to every harmonic:

$$\Delta\varphi_1^{(k)} = \Psi\left(f_0'^{(k-1)}, f_0'^{(k)}, N\right) - \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) \tag{13}$$

$$\varphi_l'^{(k)} = \varphi_l^{(k)} + l\sum_{q=1}^{k} \Delta\varphi_1^{(k)} \quad l = 1 \ldots L^{(k)} \quad \forall k \tag{14}$$

The modification factor for the pitch can be constant or time-varying. The stochastic component is not modified.

4. Concatenation of units

A database of units is built by analyzing a number of sentences uttered by the same speaker. A typical TTS front-end is used in our synthesis system: the selected units, durations, energies and pitch contours are given as input data.

The deterministic plus stochastic representation of each unit is transformed to match the energy, duration and pitch specifications, and the resulting data structures are concatenated together. The energy scaling is performed by multiplying the amplitudes and the gain of the LPC filters by the desired factor. At the border between two consecutive units, two aspects must be taken into account. The most important is the phase coherence. Let k_A and k_B be the points at the bounds of the units to be concatenated, and N the distance between them. New phase increments are necessary after the prosodic modifications to make the waveforms match properly:

$$\varphi_j'^{(k)} = \varphi_j^{(k)} + \left\{\varphi_j^{(k_A)} - \varphi_j^{(k_B)} + \Psi\left(f_0^{(k_A)}, f_0^{(k_B)}, N\right)\right\} \tag{15}$$
$$j = 1 \ldots J^{(k)} \quad k = k_B, k_B + 1, \ldots$$

The delay applied to each sinusoid is constant over k, and the relation between them may not be linear. As a result of this, the original waveform of the second unit is modified. Nevertheless, our experiments indicate that this modification is not perceived by the listener, as it is explained in section 5. Furthermore, the algorithms proposed for the prosodic modifitacions still work well. This phase adjustment is the main difference between the new system and the previously reported one [11].

The second adjustment is carried out in the amplitudes of the sinusoids near the borders between the two units, to spectrally smooth the transition:

$$A_j'^{(k)} = \eta\left[\mu A_j^{(k)} + (1-\mu) B_j^{(k)}\right] \tag{16}$$

The parameter μ is linearly increased from 0.5 to 1.0 until a certain distance from the point of concatenation is reached. The amplitudes $B_j^{(k)}$ are obtained by linear interpolation of $A_j^{(k_A)}$ measured in dB if $k > k_A$, or $A_j^{(k_B)}$ if $k < k_B$. The parameter η is used to keep the previous energy value.

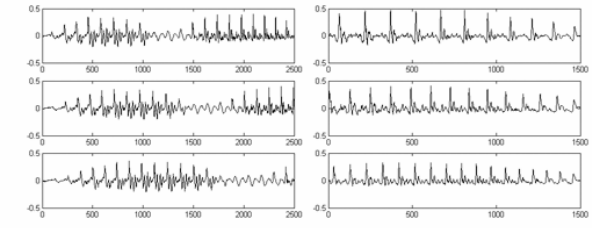

Figure 1: Duration and pitch modification. Factors: 0.8, 1.0, 1.25

5. Discussion

A database of more than 500 sentences, each of them uttered by two different speakers, was recorded at a sampling frequency of 16 KHz. The basic units were chosen to be diphones. The parameters used during the analysis were the following: N=128 samples (8 ms), f_{max}=5 KHz, 14th order LPC filters for the stochastic component. The resynthesis without modification of the analyzed data was used to validate the analysis, and the synthetic output of the system resulted to be almost indistinguishable from the original. The pitch and

duration modification of the signals were subjectively compared with the obtained by means of the PSOLA method, and their performance was found to be similar. The range of allowable modification factors was also similar, but a higher quality was reached by our system when the signal was strongly lengthened (ρ>2), because the duplication of periods in PSOLA gives to the sound a certain metallic aspect.

It was also discovered that if a constant angle α_j is added to the phase of the j^{th} harmonic at every point k, the resulting synthetic signal is perceptually equivalent to the unmodified one. Furthermore, it was found that the performance of the prosodic modifications is not affected by this kind phase increments. We have taken advantage of this fact to obtain eq. 19, which indicates how to alter the phases of the units to be concatenated. It was concluded that the relationship between the phases of the $J^{(k)}$ harmonics at a certain point k is not as important as the relative variation of each of the $J^{(k)}$ phases over k.

Our method for duration and pitch modification of the sinusoidal component is advantageous with respect to other methods found in the literature. It is much more simple than the proposed by McAulay and Quatieri [4], because it works without onset times and deals with amplitudes and phases without separating the contribution of the vocal tract from the signal. The methods proposed in [7] and [8] have been designed for a pitch-synchronous analysis/synthesis, in which the input signal has to be divided in a very precise way into its fundamental periods. On the other hand, our method can be used to modify the data independently of the time instants in which they were measured, and no duplication/deletion of pitch periods is done in the modification process. In [9] most of the mentioned problems are solved, but the application of an inverse filtering technique is required to estimate the vocal tract and then modify the pitch of the input signal. In addition, the 3rd order polynomials used in [9] to calculate the phase increments are substituted in our system by more simple operations.

The concatenative synthesis system described was implemented and several sentences were generated. The quality and the naturalness of the resulting sound were found to be very high. The same sentences were synthesized by means of the TD-PSOLA method. During the evaluation, seven different people were asked to listen to both output signals and give their opinion about the concatenation, the quality of the synthetic sound and any other differences perceived. None of the listeners was an specialist. All of them considered that our system was better in terms of concatenation, and the quality was found to be similar. One of the listeners perceived a better quality in the PSOLA signal for a low-pitched voice, below 90 Hz.

6. Conclusions

In this paper we describe a high-quality synthesis system in which the prosodic modifications are performed by means of simple and efficient operations like linear interpolations. The system is based on the deterministic plus stochastic model of speech. Pitch-synchrony is not necessary for the analysis-synthesis process. The concatenation method has been optimized to avoid spectral and waveform mismatches.

In future works, the described system will be used to develop a multilingual voice conversion tool.

7. Acknowledgements

This work was partially supported by TC-STAR, FP6-506738.

8. References

[1] Dutoit, T. "An Introduction to Text-to-Speech Synthesis", Kluwer Academic Publishers, chapt.10, 1997.

[2] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, J. Schroeter, "TD-PSOLA versus Harmonic plus Noise Model in Diphone based Speech Synthesis", Proc. ICASSP 98, pp. 273-276, 1998.

[3] R.J. McAulay, T.F. Quatieri, "Speech Analysis/Synthesis based on a Sinusoidal Representation", IEEE Trans. on Acoust., Speech and Signal Processing, Vol. 34 (4), pp. 744-754, 1986.

[4] T.F. Quatieri, R.J. McAulay, "Shape Invariant Time-Scale and Pitch Modification of Speech", IEEE Transactions on Signal Processing, Vol. 40 (3), pp. 497-510, 1992.

[5] X. Serra, "A System for Sound Analysis/Transformation/ Synthesis Based on a Deterministic plus Stochastic Decomposition", PhD. thesis, Stanford University, 1989.

[6] J. Laroche, Y. Stylianou, E. Moulines, "HNM: a Simple, Efficient Harmonic + Noise Model for Speech", Proc. ICASSP 93, 1993.

[7] E.R. Banga, C. García Mateo, X. Fernández Salgado, "Concatenative Text-to-Speech Synthesis based on Sinusoidal Modeling", Improvements in Speech Synthesis, ISBN 0471 49985 4, pp. 39-51, John Wiley and Sons, 2001.

[8] Y. Stylianou, "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification", PhD. Thesis, 1996.

[9] D. O'Brien, A. I. C. Monaghan, "Concatenative Synthesis based on a Harmonic Model", IEEE Trans. on Speech and Audio Processing, Vol. 9 (1), pp. 11-20, 2001.

[10] M.W. Macon, M.A. Clements, "Speech Concatenation and Synthesis using an Overlap-Add Sinusoidal Model", Proc. ICASSP 96, pp. 361-364, 1996.

[11] D.Erro, A.Moreno, "A Pitch-Asynchronous Simple Method for Speech Synthesis by Diphone Concatenation using the Deterministic plus Stochastic Model", Proc. SPECOM, 2005.

[12] Ph. Depalle, T. Hélie, "Extraction of Spectral Peak Parameters using a STFT Modeling and no Sidelobe Windows", Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1997.