Flexible Harmonic/Stochastic Speech Synthesis

Daniel Erro, Asunción Moreno, Antonio Bonafonte

TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
{derro,asuncion,antonio}@gps.tsc.upc.edu

Abstract

In this paper, our flexible harmonic/stochastic waveform generator for a speech synthesis system is presented. The speech is modeled as the superposition of two components: a harmonic component and a stochastic or aperiodic component. The purpose of this representation is to provide a framework with maximum flexibility for all kind of speech transformations. In contrast to other similar systems found in the literature, like HNM, our system can operate using constant frame rate instead of a pitch-synchronous scheme. Thus, the analysis process is simplified, while the phase coherence is guaranteed by the new prosodic modification and concatenation procedures that have been designed for this scheme. As the system was created for voice conversion applications, in this work, as a previous step, we validate its performance in a speech synthesis context by comparing it to the well-known TD-PSOLA technique, using four different voices and different synthesis database sizes. The opinions of the listeners indicate that the methods and algorithms described are preferred rather than PSOLA, and thus are suitable for high-quality speech synthesis and for further voice transformations.

1. Introduction

In concatenative speech synthesis, a set of recorded speech units are selected from a database and are concatenated to create synthetic utterances. The prosodic characteristics of the units are adapted to the desired prosodic contour and the discontinuities between the different units are minimized at the boundaries. The performance of the speech synthesis systems strongly depends on the techniques and algorithms used for all these tasks. Furthermore, voice conversion methods are usually integrated into speech synthesis systems as a complement used to modify the physical attributes of the output voice to be perceived by the listeners like a different voice. This fact makes desirable the choice of flexible signal models capable of providing a high degree of flexibility without causing artifacts.

In [1] we presented a new simple method for prosodic modification of speech and for concatenation of speech units. The harmonic plus stochastic model of speech (HSM) was used to implement the waveform generation block of a text-to-speech synthesis system (TTS), due to the flexibility and capacity of manipulation provided by the model, as well as its interesting properties for embedded systems. Unlike in Stylianou's HNM [2] and other similar methods, the prosodic modifications were not based on pitch-synchronous overlap-add (PSOLA) techniques [3]. The main advantage of the system was that, although neither pitch marks nor accurate separation of signal periods were necessary, the inter-frame phase coherence and the speech waveform shape invariance were successfully maintained by means of new phase manipulation algorithms. Therefore, the analysis of speech was simplified using a constant frame rate, whereas the usage of onset times, source-filter separation techniques and cross-correlation-based phase corrections was also avoided, in contrast to other previous non-pitch-synchronous sinusoidal systems [4, 5, 6]. Instead, the modification algorithms designed for the new method were conceptually simple and straightforward.

At present, successful voice conversion methods compatible with HSM have been designed and tested on natural speech in a public evaluation campaign in both, intra-lingual and cross-lingual applications, achieving excellent results [13]. In this paper we describe the full HSM-based waveform generation block, which has been integrated into a TTS system. New improved phase manipulation algorithms, related to the prosodic modification and concatenation of speech units, are explained in detail. The purpose of the comparative experiments conducted in this paper is to validate the suitability of our waveform generator for high-quality speech synthesis, prior to using it for converting synthetic voices. A brief explanation about our voice conversion method is also included, although it does not take part in the discussion here.

The paper is structured as follows. Section 2 shows how the speech signals are analyzed and reconstructed from the measured parameters. Section 3 describes the algorithms for modifying the signal parameters in order to change the pitch or duration of speech. Section 4 deals with the artifact-free concatenation of speech units. The system is extended with the WTW voice conversion method in section 5. In section 6 the performance of the speech synthesis system is evaluated by comparing it to the UPC TTS system Ogmios [7]. The main conclusions of this work are listed in section 7.

2. Analysis and Reconstruction of Signals

The harmonic plus stochastic model (HSM) [2] assumes that the speech signal can be represented as a sum of a number of harmonically related sinusoids with time-varying parameters and a noise-like component. The harmonic component is present only in the voiced fragments of speech. It can be represented at each analysis frame by the fundamental frequency and the amplitudes and phases of the harmonics. The stochastic component tries to model all the non-sinusoidal signal components, caused by the frication,

breathing noise, etc. It can be represented at each frame by the coefficients of an all-pole filter.

2.1. Analysis

The signals are analyzed using a constant frame rate of 100 frames per second. Given a speech frame to be analyzed, frame number k, the fundamental frequency $f_0^{(k)}$ has to be estimated and a binary voicing decision is taken. If the frame is voiced, the amplitudes $\{A_j^{(k)}\}$ and phases $\{\varphi_j^{(k)}\}$ of all the harmonics below a cutoff frequency of 5 KHz are detected. The choice of a fixed cutoff frequency is adequate for voice conversion purposes, because the spectral envelopes are extracted from the harmonic component. The amplitudes and phases are obtained by means of a least squares optimization in the spectral domain, using the algorithm of Depalle et al. [8] particularized to harmonic sinusoids:

$$S^{(k)}(f) = \sum_{j=1}^{J^{(k)}} \tfrac{1}{2} A_j^{(k)} \left[e^{i\varphi_j^{(k)}} W(f - jf_0) + e^{-i\varphi_j^{(k)}} W(f + jf_0) \right] \quad (1)$$

$S^{(k)}(f)$ is the STFT of the k^{th} frame and $W(f)$ denotes the Fourier transform of the analysis window, whose length is two pitch periods. $J^{(k)}$ is the highest integer that satisfies $J^{(k)} \cdot f_0^{(k)} < 5$ KHz. As the optimization is performed in the spectral domain, the relative position of the analysis window within the pitch period is not important. This is adequate for a pitch-asynchronous analysis framework.

Once the frequencies, amplitudes and phases of the harmonics are known, the sinusoidal component of the signal is regenerated by interpolating between the measured values. For each time instant, the instantaneous amplitudes are obtained by means of a linear interpolation, and the 3rd order polynomial proposed by McAulay and Quatieri [9] is used to interpolate the instantaneous frequencies and phases of each harmonic. The regenerated harmonic component is subtracted from the original signal, and the remaining part of the signal, which corresponds to the stochastic component, is LPC-analyzed at each frame.

2.2. Reconstruction

The signal is reconstructed by overlapping and adding $2N$-length frames, where N is the distance between the analysis frame centres, measured in samples. Each synthetic frame contains the sum of the measured harmonics with constant amplitudes, frequencies and phases, and the stochastic contribution, generated by filtering white gaussian noise with the measured LPC-filters. A triangular window is used to overlap-add the frames in order to obtain the time-varying synthetic signal. Let k be the frame number and j the harmonic number. The following expressions are used to reconstruct the signal.

$$s^{(k)}[n] = \sum_{j=1}^{J^{(k)}} A_j^{(k)} \cos\left(2\pi j f_0^{(k)} n / f_s + \varphi_j^{(k)}\right) + \sigma[n] * h_{LPC}^{(k)}[n] \quad (2a)$$

$$s[kN + m] = \left(\tfrac{N-m}{N}\right) \cdot s^{(k)}[m] + \left(\tfrac{m}{N}\right) \cdot s^{(k+1)}[m - N] \quad (2b)$$

where m is in the range [0, N–1]. The speech signal resynthesized from the measured parameters is almost indistinguishable from the original.

3. Prosodic Modifications

As a pitch-asynchronous scheme is being used, the prosodic modification of the signal implies the challenge of modifying the phases of the harmonics without altering the phase coherence between frames or causing artifacts. For this purpose, we have developed new strategies to manipulate the phases. We consider that the phases $\varphi_j^{(k)}$ measured at a certain analysis frame k are the sum of two components: a linear-in-frequency term given by the parameter $\alpha^{(k)}$, and the phase contribution of the time-varying vocal tract, $\theta_j^{(k)}$.

$$\varphi_j^{(k)} = j\alpha^{(k)} + \theta_j^{(k)} \quad (3)$$

The estimation of $\alpha^{(k)}$ is discussed in section 3.3.

3.1. Duration Modification

The duration modification can be carried out by increasing or decreasing the distance N between the synthesis points in equation (2b), so that the amplitude and fundamental frequency variations get adapted to the new time scale. On the other hand, if the phases were kept unmodified, fixed at the center of the frames, the waveform coherence between consecutive points would be lost, causing artifacts and noisy pitch variations. Therefore, the change in N needs to be compensated with a phase manipulation in a way that the waveform and pitch of the duration-modified signal are similar to the original. This manipulation should affect only to the linear-in-frequency phase term. Assuming that the fundamental frequency varies linearly from frame k–1 to k, we define the function Ψ which represents the expected phase increment of the first harmonic between those points, affecting only the linear-in-frequency term:

$$\alpha^{(k)} - \alpha^{(k-1)} \cong \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) = \pi N \left(f_0^{(k-1)} + f_0^{(k)}\right) / f_s \quad (4)$$

If N is substituted by N', the following phase correction is applied:

$$\Delta\varphi_1^{(k)} = \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N'\right) - \Psi\left(f_0^{(k-1)}, f_0^{(k)}, N\right) \quad (5a)$$

$$\varphi_j'^{(k)} = \varphi_j^{(k)} + j\sum_{\kappa=2}^{k} \Delta\varphi_1^{(\kappa)} \quad j = 1 \ldots J^{(k)} \quad \forall k > 1 \quad (5b)$$

This correction compensates the modification of N without affecting the small local variations in the vocal tract phase response. The stochastic coefficients are not modified. Note that the modification factor can be time-varying.

3.2. Pitch Modification

For the pitch modifications, the amplitudes of the new harmonics $A_j'^{(k)}$ are obtained by a simple linear interpolation between the measured log-amplitudes in order to maintain the formant structure unaltered. A constant multiplicative factor is used to keep constant the energy of the harmonic component despite the variation of the number of sinusoids. The vocal tract contribution to the phases of the new harmonics, $\theta_j'^{(k)}$, can be obtained by means of a linear interpolation of the real and imaginary parts of the complex amplitudes $A_j^{(k)}\exp(i\cdot\theta_j^{(k)})$. The values of $\theta_j^{(k)}$ are calculated from the original phases $\varphi_j^{(k)}$ by subtracting the linear-in-frequency phase term given by $\alpha^{(k)}$.

$$\theta_j^{(k)} = \varphi_j^{(k)} - j\alpha^{(k)} \quad (6)$$

Finally, the relative position of the synthesis point within the new pitch period is now different and the linear term has to be corrected to compensate the modification of the periodicity. The phase correction to be performed is given by (5b) with

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

The stochastic coefficients are not modified. Time-varying modification factors can be used following this method, and the simultaneous duration + pitch modification of the signal is also possible.

3.3. Linear Phase Term Estimation

The estimation of the parameter $\alpha^{(k)}$ is a crucial point for obtaining high quality synthetic speech. In pitch synchronous systems the linear phase term is zero at the frame centres, but in exchange a set of pitch marks need to be stored and synchronized with the waveform. Even in such systems, some problems appear when there are linear phase mismatches between different signal periods [10]. We propose to estimate $\alpha^{(k)}$ using the following formula.

$$\alpha^{(k)} = \arg\max_{\alpha} \sum_{j=1}^{J^{(k)}} A_j^{(k)} \cos\left(\varphi_j^{(k)} - j\alpha\right) \quad (8)$$

Thus, the linear phase term is considered to be zero near the maximum of the waveform defined by the measured harmonics. Note that this strategy is similar to the one followed in some pitch-synchronous systems in which the two-period-length frames are separated using the signal maxima as reference. The underlying assumption is that the waveform reaches its maximum when the phases of the harmonics are maximally close to zero. The maximum of the summation is one of the zeros of its derivative:

$$\sum_{j=1}^{J^{(k)}} \left[jA_j^{(k)} \sin\varphi_j^{(k)} \cos(j\alpha) - jA_j^{(k)} \cos\varphi_j^{(k)} \sin(j\alpha) \right] = 0 \quad (9)$$

The resulting equation is nonlinear, but it can be simplified using the substitution $x=\cos\alpha$ and the Tsebyshev polynomials, defined recursively as:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x) \quad (10a)$$

$$U_0(x) = 1, \quad U_1(x) = 2x, \quad U_n(x) = xU_{n-1}(x) + T_n(x) \quad (10b)$$

These polynomials verify the following conditions:

$$\cos j\alpha = T_j(x) \quad (11a)$$

$$\sin j\alpha = \sin\alpha \cdot U_{j-1}(x) = \pm\sqrt{1-x^2} \cdot U_{j-1}(x) \quad (11b)$$

so equation (9) can be transformed into

$$P(x) \pm \sqrt{1-x^2}\,Q(x) = 0 \quad (12)$$

where P and Q contain the weighted sum of T-type and U-type polynomials, respectively. The solutions of (12) are also solutions of

$$P(x)^2 - (1-x^2)Q(x)^2 = 0 \quad (13)$$

Among all the solutions of (13), which are easily located by any typical root finding method between $x=-1$ and $x=1$, the one that maximizes (8) is chosen and its corresponding α is calculated.

In practice, not all harmonics need to be used for the calculation of the linear phase term. Only the most powerful harmonics are relevant for this task, so the complexity of the problem can be reduced by selecting only those harmonics.

It must be taken into account that the polarity of the signals is not always the same. This algorithm is designed for signals in which the positive peaks are greater than the negative peaks. In the other case, equation (8) should be minimized instead of maximized.

4. Concatenation of Units

In concatenative speech synthesis, the synthetic utterances are built by concatenating different speech units selected from a recorded database. The prosodic contour of the units is adapted to the desired specifications, given by a prosody generation block whose input is the text to be pronounced by the system. The algorithm for concatenation of units recorded in different phonetic contexts has to minimize the waveform discontinuities and the spectral mismatches at the boundaries.

In order to develop a waveform generation block using the HSM, the whole database has to be analyzed and parameterized according to the model. Once the prosody of the selected units is modified, the waveform discontinuities are avoided by correcting the linear phase term of the incoming unit to be coherent with the previously concatenated units. Let k^A be the last frame of the last unit concatenated A, and k^B the first frame of the incoming unit B. The phase correction is given by the following expressions:

$$\Delta\varphi_1^{AB} = \alpha^{(k_A)} - \alpha^{(k_B)} + \Psi\left(f_0^{(k_A)}, f_0^{(k_B)}, N\right) \quad (14a)$$

$$\varphi_j'^{(k)} = \varphi_j^{(k)} + j\Delta\varphi_1^{AB}, \quad k \geq k_B \quad (14b)$$

where α is calculated using equation (8). It must be emphasized that in the case of concatenative synthesis the calculation of the linear phase terms is performed only once, when building the synthesis database, so that α is stored together with the rest of signal parameters.

On the other hand, a smoothing technique is applied to the amplitude envelopes of the frames near the unit boundaries, so that the spectral discontinuities are also minimized.

5. Voice Conversion by WFW

In general, voice conversion systems apply a previously trained transformation function to the input signal. In our case, the input signals are synthetic utterances obtained by concatenation of selected units. Thus, the TTS system acts as source speaker. In our system, the Weighted Frequency Warping method (WTW), recently proposed by the author [11], is used for voice conversion. This method has been already tested with natural speech, and the results show that a good balance between quality and conversion degree is obtained. In the framework of the TC-STAR project, our voice conversion system was evaluated in both intra-lingual and cross-lingual contexts, and excellent results were obtained [13]. Although the WFW method is not discussed or evaluated in this paper, the voice conversion algorithm is described in this section in order to offer complete information about the waveform generation process.

5.1. Prosodic Conversion

During the training phase, the mean μ and standard deviation σ of the $\log f_0$ are determined for the source and target speakers. During the conversion phase, given a synthetic utterance generated by the TTS system, the pitch

contour is modified to match the specifications of the target speaker according to the following expression:

$$\log f_0^{(\text{converted})} = \mu^{(\text{target})} + \frac{\sigma^{(\text{target})}}{\sigma^{(\text{source})}}\left(\log f_0^{(\text{source})} - \mu^{(\text{source})}\right) \quad (15)$$

5.2. Spectral Conversion

The spectral transformation concerns the amplitudes and phases of the harmonics and the LPC coefficients of the stochastic component. During the training phase, a gaussian mixture model (GMM) of m gaussian components is trained from a set of phonetically aligned source-target acoustic vector pairs $\{[x^T\ y^T]^T\}$ [12]. The joint source-target GMM is represented by the weights $\{\alpha_i\}$, the mean vectors $\{\mu_i\}$ and the covariance matrices $\{\Sigma_i\}$ of each of the gaussian components. In this work, the training vectors $\{x\}$ and $\{y\}$ contain the line spectral frequencies (LSF) that represent the all-pole filter that better fits the amplitudes of the harmonics. Once the GMM has been trained, given a source LSF vector x, the probability that x belongs to the i^{th} gaussian component of the model, $p_i(x)$, is given by

$$p_i(x) = \frac{\alpha_i N\left(x, \mu_i^x, \Sigma_i^{xx}\right)}{\sum_{j=1}^{m} \alpha_j N\left(x, \mu_j^x, \Sigma_j^{xx}\right)} \quad (16)$$

where μ_i^x and Σ_i^{xx} can be extracted from μ_i and Σ_i, respectively.

$$\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \quad \Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \quad (17a, b)$$

Using the information provided by the GMM, a different frequency warping function $W_i(f)$ is calculated for each gaussian component i between μ_i^x and μ_i^y. As they both are LSF vectors, the formants given by their corresponding all-pole filters are used as reference points for a piecewise linear frequency warping function. Finally, to conclude with the training procedure, a new function is designed to predict the stochastic component of the target speaker from the LSF representation of its harmonic component. The stochastic LPC coefficients associated with each of the training LSF vectors $\{y\}$ are also translated into LSF vectors $\{y_{st}\}$, and matrices $\{\Gamma_i\}$ and vectors $\{v_i\}$ are found so that the following prediction function is optimized:

$$y_{st} = \sum_{i=1}^{m} p_i(y) \cdot \left[\eta_i + \Gamma_i \left(\Sigma_i^{yy}\right)^{-1} \left(y - \mu_i^y\right)\right] \quad (18)$$

where μ_i^y and Σ_i^{yy} are used in equation (16) to obtain $p_i(y)$. At the end of the training phase, the GMM parameters, the frequency warping functions and the stochastic prediction function have been calculated.

In the conversion phase, given a source frame to be converted, the associated LSF vector x is extracted from the amplitudes of the harmonics, and the m probabilities $p_i(x)$ are calculated using expression (16). The individual warping function of the current frame is obtained as a linear combination between the m trained basis functions $W_i(f)$.

$$W(f) = \sum_{i=1}^{m} p_i(x) \cdot W_i(f) \quad (19)$$

We assume that phonemes with similar formant structures, which are linked to the same gaussian component of the GMM, should be associated with similar frequency warping trajectories. Thus, the probabilities $p_i(x)$ are used as weights for the linear combination of the m different warping trajectories. The magnitude envelope $A(f)$ of the current frame is estimated by means of a linear interpolation between the measured harmonic log-amplitudes. The phase envelope $\theta(f)$ is estimated by linearly interpolating the real and imaginary parts of the complex amplitudes $A_j^{(k)}\exp(i\cdot\theta_j^{(k)})$, as in section 3.2. Warped envelopes $A'(f)$ and $\theta'(f)$ are calculated, and the target amplitudes $\{A_j'^{(k)}\}$ and vocal tract phases $\{\theta_j'^{(k)}\}$ are calculated by resampling them at the positions of the harmonics.

$$A'(f) = A\left(W^{-1}(f)\right), \quad \theta'(f) = \theta\left(W^{-1}(f)\right) \quad (20a, b)$$

This step does not completely transform the source voice into the target speaker's voice because the formants are only reallocated while their amplitude remains unmodified. Therefore, the energy distribution is corrected using the converted LSF vector $F(x)$, which is obtained by means of the typical GMM-based transformation function:

$$F(x) = \sum_{i=1}^{m} p_i(x) \cdot \left[\mu_i^y + \Sigma_i^{yx}\left(\Sigma_i^{xx}\right)^{-1}\left(x - \mu_i^x\right)\right] \quad (21)$$

The energy of the envelope given by $F(x)$ is measured at the bands 100-300Hz, 300-800Hz, 800-2500Hz, 2500-3500Hz and 3500-5000Hz, which are likely to contain different formants. Multiplicative factors are used inside each band to correct the energy of the frequency-warped harmonics. Finally, the stochastic component of the converted frame is predicted using expression (18), in which y is substituted by the converted LSF vector $F(x)$ (21). The stochastic component of the unvoiced frames is left unmodified, because its conversion does not lead to any important improvement and it can cause a small loss of quality.

6. Experiments and Discussion

A preference test was carried out in order to determine if the proposed algorithms were suitable for the development of a high quality speech synthesis system. Ogmios is the speech synthesis system that has been created at the UPC [7]. It is based on unit selection, and it includes a waveform generation block based on the TD-PSOLA technique, which cannot be used for voice conversion but is almost standard for synthesis. For the preference test, the text processing, prosody generation and unit selection blocks of Ogmios were used to obtain the sequence of units and prosodic specifications of the different synthetic utterances, and the audio samples were generated using both Ogmios and a new waveform generation block based on the HSM and the algorithms described in the previous sections.

In order to emphasize the effectiveness of both methods in speech modification and concatenation, the system was forced to modify the prosody of all the selected units to match the specifications provided by the prosody generation block of Ogmios. Under these conditions, the artifacts introduced by both methods were more visible for the comparison, although the quality of the sentences was lower.

The 18 listeners that participated in the test, 6 speech synthesis experts and 12 volunteers, were asked to listen to 17 pairs of synthetic utterances in Spanish. All the listeners were native Spanish speakers. Four different voices were used in

this experiment. Two of them, one male and one female, were built from a database consisting of more than 10 hours of recorded speech. The databases of the two remaining voices, male and female, contained less than an hour of recorded speech. 10 of the sentence pairs in the test were generated from the large-database voices, and 7 pairs were built from the small-database voices. For each sentence pair, whose components were played in random order, the listeners were asked to choose between the following options: "I prefer the first", "I prefer the second" or "I can't decide". The results of the preference test are shown in figure 1.

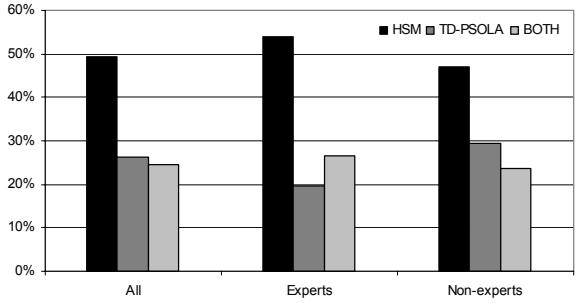

Figure 1: results of the preference test.

Figure 2 shows separately the results for large synthesis databases (a) and for small synthesis databases (b). In figure 3 individual results for female voices (a) and for male voices (b) are displayed separately.

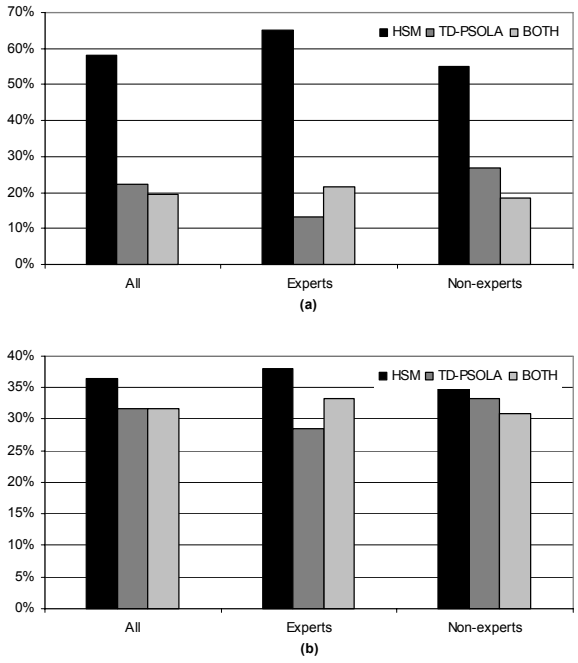

Figure 2: results for large (a) and small databases (b).

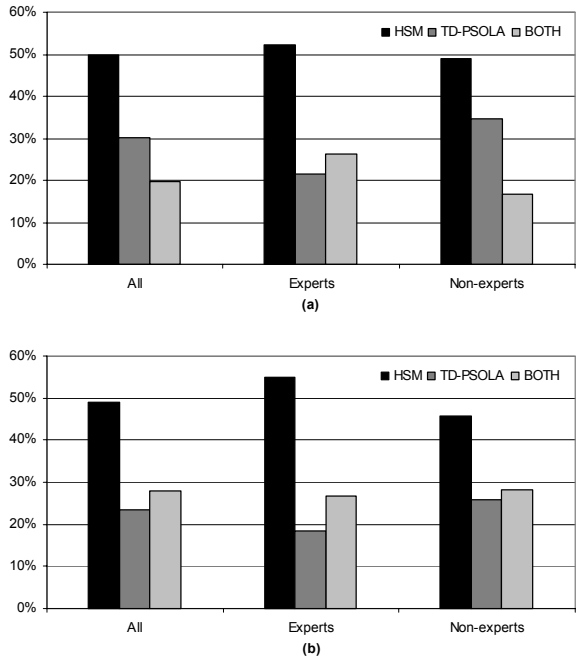

Figure 3: results for female (a) and male (b) voices.

As it can be seen, in the conditions of this experiment the new HSM waveform generation block clearly outperforms the one based on TD-PSOLA. This assertion holds for both expert and non-expert listeners, but the new method is slightly better scored by experts. Concerning figure 2, it can be observed that when the synthesis databases are small, the uncertainty increases and the scores are closer to each other. This fact can be a consequence of the different noise sources in each case. When the databases are large, all the phonemes are represented by a high number of instances. Thus, the prosodic modification factors needed are lower and the associated noise is less important than the artifacts coming from the concatenation of units. The concatenations obtained by means of the HSM algorithms are smoother because the spectral envelopes can be manipulated. On the contrary, when the synthesis database is small, the loss of quality caused by the prosodic modifications and by severe concatenation artifacts affects both methods in a more similar way. Figure 3 shows that the scores reached by the HSM waveform generator are similar in both genders.

The experiment described shows that the HSM method and the algorithms presented in this paper, which have been successfully used for voice conversion purposes, are also suitable for high-quality speech synthesis without voice conversion. The listeners' choices seem to be more influenced by the concatenation properties than by the quality of the prosodic modification. However, the results may be different for other implementations of the unit selection procedure that assign a lower weight to the prosodic aspects of the units and a higher weight to the spectral aspects. In addition, it must be taken into account that in a standard synthesis application not all the units are prosodically modified, and in this situation the TD-PSOLA approach can be expected to reach higher scores because it works directly with the recorded speech samples.

7. Conclusions

In this paper we have presented our improved waveform generation system for speech synthesis based on the harmonic plus stochastic model. The new algorithms for prosodic modification, concatenation and conversion of speech, which in contrast to other methods do not require pitch-synchronism, have been described in detail. The experiments carried out in this paper show that the listeners prefer this new approach to a more standard TD-PSOLA approach. It can be concluded that the algorithms and methods described, which were successfully used for voice conversion applications, are also suitable for high-quality speech synthesis.

In future works voice conversion constraints will be included in the cost function of the unit selection block of a TTS system. It is expected that the performance of the synthesis + conversion system will be improved if the units that are easier to convert are assigned a higher probability to be selected for synthesis.

8. Acknowledgements

This work was partially supported by TC-STAR (Technology and Corpora for Speech-to-Speech Translation, FP6-506738) and AVIVAVOZ (TEC2006-13694-C03).

9. References

[1] Erro, D., Moreno, A., "A pitch-asynchronous simple method for speech synthesis by diphone concatenation using the deterministic plus stochastic model", *Proc. 10th Int. Conf. on Speech and Computer*, pp.321-324, 2005.

[2] Stylianou, Y., "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification", PhD thesis, École Nationale Supérieure des Télécommunications, 1996.

[3] Moulines, E., Charpentier, F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, vol.9 no.5-6 pp.453-467, 1990.

[4] Quatieri, T.F., McAulay, R.J., "Shape invariant time-scale and pitch modification of speech", *IEEE Transactions on Signal Processing*, 1992.

[5] O'Brien, D., Monaghan, A.I.C., "Concatenative synthesis based on a harmonic model", *IEEE Transactions on Speech and Audio Processing*, 2001.

[6] Chazan, D., Hoory, R., Sagi, A., Shechtman, S., Sorin, A., Shuang, Z.W., Bakis, R., "High quality sinusoidal modeling of wideband speech for the purposes of speech synthesis and modification", *ICASSP*, 2006.

[7] Bonafonte, A., Agüero, P.D., Adell, J., Pérez, J., Moreno, A., "OGMIOS: The UPC text-to-speech synthesis system for spoken translation", *TC-Star Workshop on Speech to Speech Translation*, 2006.

[8] Depalle, Ph., Hélie, T., "Extraction of spectral peak parameters using a STFT modeling and no sidelobe windows", *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[9] McAulay, R.J., Quatieri, T.F., "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1986.

[10] Stylianou, Y., "Removing linear phase mismatches in concatenative speech synthesis", *IEEE Transactions on Speech and Audio Processing*, 2001.

[11] Erro, D., Moreno, A., "Weighted Frequency Warping for Voice Conversion", *InterSpeech*, 2007.

[12] Kain, A., "High resolution voice transformation", PhD thesis, OGI School of Science and Engineering, 2001.

[13] Choukry, K., et al. Evaluation Report. Deliverable D30 of the EU funded project TC-STAR. March 2007.