AhoTransf: A tool for Multiband Excitation based speech analysis and modification

Ibon Saratxaga, Inmaculada Hernáez, Eva Navas, Iñaki Sainz, Iker Luengo, Jon Sánchez, Igor Odriozola, Daniel Erro

Aholab - Dept. of Electronics and Telecommunications. Faculty of Engineering. University of the Basque Country

Urkijo zum. z/g 48013 Bilbo

{ibon, inma, eva, inaki, ikerl, ion, igor, derro}@aholab.ehu.es

Abstract

In this paper we present AhoTransf, a tool that enables analysis, visualization, modification and synthesis of speech. AhoTransf integrates a speech signal analysis model with a graphical user interface to allow visualization and modification of the parameters of the model. The synthesis capability allows hearing the modified signal thus providing a quick way to understand the perceptual effect of the changes in the parameters of the model. The speech analysis/synthesis algorithm is based in the Multiband Excitation technique, but uses a novel phase information representation the Relative Phase Shift (RPS's). With this representation, not only the amplitudes but also the phases of the harmonic components of the speech signal reveal their structured patterns in the visualization tool. AhoTransf is modularly conceived so that it can be used with different harmonic speech models.

1. Introduction

Speech models based in the separation of the periodic and the noise-like parts of the speech were early introduced in the speech processing panorama. The early work by McAulay and Quatiery (1986) with the sinusoidal modelling, where the signal was modelled by means of sinusoidal components located at the frequencies where the peaks of the spectrum were, was quickly followed by the harmonic systems (Griffin & Lim, 1988; Laroche, Stylianou, & Moulines, 1993; Stylianou, 1996). This harmonic constraint is appropriate for the speech signal and simplified the analysis and the synthesis, eliminating the need of peak picking and peak tracking algorithms.

However, modelling only the harmonic part of the signal leaves out quite a lot of information, so harmonic models were complemented with a noise-like component. This noisy component has been defined in different ways: some proposals (Laroche, Stylianou & Moulines 1993; Stylianou, 1996) assume that the noise is above a certain frequency (harmonic plus noise family, HNM); others overlap the harmonic and the noise-like parts along part or all of the spectrum (Stylianou, 1996; Erro, Moreno & Bonafonte, 2007) (harmonic plus stochastic family); and finally others interleave periodic and noisy components in harmonic bands, (Griffin & Lim, 1988; Dutoit & Leich, 1993) (multiband excitation family, MBE). The model implemented in the tool described in this paper falls into this last category.

When these models were first proposed (late eighties and early nineties) they meant an important leap towards voice quality, because they allowed high quality coding and thus good synthetic voice quality. Being fully parametric, they solved the problem of concatenation mismatches and allowed easy pitch and duration modifications of the signals. They also permitted low bit rate high quality coding. The main downside was their complexity and the heavy computational requirements of the analysis stage. The arrival of the unit selection techniques for synthesis, which produced higher naturalness and required comparably less computational effort, slowed down the development of these methods.

Nevertheless, more recently HNM models have gained more and more interest, as more and more research effort is being oriented towards the area of voice transformation and voice conversion. Sure enough, the parametric nature of these models allows not only pitch and duration transformations but also spectral manipulation, and it has been reported that strong modifications can be done to the signal while keeping a certain degree of naturalness (Stylianou, 1996).

Our interest in this area derives also from its application to voice transformation in general, and we have developed several HNM models, seeking the higher possible level of naturalness for speech. We have built up a Harmonic plus Noise model based on the Multiband Excitation techniques, but with specific phase control techniques (Saratxaga et al., 2009) developed by us. The resulting system appears suitable for voice transformation: it is robust, it is pitch asynchronous, it has good quality, it is fully parametric, and the parameters are quite straightforward, so as they can be easily manipulated.

To gain a better understanding of the relationship between the parameters of such a model and the perceptual characteristics of the speech we have developed the AhoTransf tool. This tool shows the different parameters of the model in spectrogram-like displays and allows modifying any of these parameters. It integrates a re-synthesis algorithm so the user can hear the effect of the modifications.

In the next section, the model is outlined in three parts: one describing the analysis stage, another the synthesis

one and the last one explaining pitch and duration modifications. Then, the functionality of AhoTransf is described in detail and finally, a conclusion section closes the paper.

2. HNM-MBE model

The proposed harmonic plus noise multiband excitation model (HNM-MBE) is based in the vocoder developed by Griffin and Lim (1988), with several modifications related to the analysis and representation of the phase of the harmonics. In this model the speech signal is decomposed into two components, a harmonic one h(t) and a noisy one n(t):

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

The MBE model considers that the whole spectrum is divided into equally wide bands centred around the pitch harmonic frequencies and each of these bands is classified as harmonic or noisy, depending on the Power Spectral Distribution (PSD) of the signal within the band. In this way, we get two components, harmonic and noisy, each of them having energy in different but interspersed frequency bands. The modelled signal can be expressed by:

$$\hat{s}(t) = \sum_{k=1}^{K(t)} \left\langle h_k(t) \,|\, n_k(t) \right\rangle \tag{2}$$

Where k denotes the band, $K(t)$ is the total number of bands at time t (which depends on the pitch value at that moment) and $h_k(t)$ and $n_k(t)$ stand for the harmonic and noisy models of the k-th band. <A|B> operator (A or B) implies a selection between the two arguments.

The harmonic bands are modelled by means of a sinusoid at the harmonic frequency, while noisy bands are modelled by a band-pass white noise. The harmonic part can thus be written as:

$$h(t) = \sum_{k=1}^{N} A_k \cos(\varphi_k) = \sum_{k=1}^{N} A_k \cos(2\pi k f_o t + \theta_k) \tag{3}$$

where N is the number of bands, the A_k are the amplitudes of the spectral envelope, φ_k is the instantaneous phase, f_o is the pitch or fundamental frequency and θ_k is the initial phase of the sinusoid.

The noise-like part can be better defined in the frequency domain, where its banded structure is clearly exposed:

$$\begin{aligned} N(\omega) &= \sum_{k=1}^{N} B_k \prod\left(\frac{\omega - 2\pi k f_0}{BW}\right) \cdot W(\omega) \quad \Big| \quad \omega \geq 0 \\ N(\omega) &= N^*(-\omega) \qquad\qquad\qquad\qquad\quad \Big| \quad \omega < 0 \end{aligned} \tag{4}$$

where B_k are the amplitudes of the noise spectral envelope in each band, BW is the bandwidth of a band and $W(\omega)$ is the Fourier transform of a sufficiently long white noise signal fragment.

2.1 HNM-MBE analysis

The analysis starts with the calculation of the fundamental frequency. A cepstrum-based pitch determination algorithm (CDP) is used for that purpose (Luengo et al., 2007). The analysis is pitch asynchronous so the frame rate can be freely chosen (8-10ms). The speech signal is windowed by means of a Hann window. The window is three pitch periods long, so as to assure a good resolution in the frequency domain where the analysis will be done.

The MBE model assumes that the spectrum of the speech signal is divided into bands centred on the pitch and its harmonics. The power spectrum is represented by an envelope with one value per band, and two of these envelopes are calculated for every analysis frame: one using the harmonic model and the other using the noise model.

2.1.1. Spectral envelopes calculation

The values of the amplitudes in every band are calculated by minimizing the energy of the modelling error of the windowed frame (Griffin & Lim, 1988):

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left| S_w(\omega) - \hat{S}_w(\omega) \right|^2 d\omega \tag{5}$$

where S_w is the windowed frame of the signal, and $\hat{S}_w$ the corresponding modelled synthetic signal.

This error is minimized when the coefficients are:

$$A_k = \frac{\int_{a_k}^{b_k} \left| S_w(\omega) \right| \left| E_w(\omega) \right| d\omega}{\int_{a_k}^{b_k} \left| E_w(\omega) \right|^2 d\omega} \tag{6}$$

where a_k and b_k are the lower and upper limits of each frequency band, and E_w is the Fourier transform of the windowed synthetic excitation signal: sum of harmonic sinusoids in the case of the harmonic model, and normalized white noise in the case of the noise model.

For the harmonic model, the Fourier Transform of a synthetic windowed excitation signal $E_w(\omega)$ is obtained for each frame.

$$E_w(\omega) = \mathcal{F}\left\{ w_{han}(t) \cdot \sum_{k=1}^{N} \cos(2\pi k f_o t) \right\} \tag{7}$$

where $w_{han}(t)$ is the aforementioned Hann window.

The Fourier transform of the signal frame, $S_w(\omega)$, is also computed and the coefficients are calculated for every band. It is worth noting that the coefficients A_k are real numbers. No complex calculation is done in this analysis. The phase of the sinusoidal components will be obtained otherwise, as it is explained in the next section.

For the noise model, the expression used to calculate the envelopes is the same as (6), but the synthetic excitation signal is much simpler: the Fourier transform of the windowed normalized white Gaussian noise equals one across the bands. Therefore, expression (6) becomes:

$$B_k = \frac{\int_{a_k}^{b_k} \left| S_w(\omega) \right| d\omega}{\int_{a_k}^{b_k} d\omega} \tag{8}$$

2.1.2. Phase calculation

Unlike the traditional MBE model, where the instantaneous phases of the harmonic components are obtained resolving a complex version of equation (6), in our model these phases are extracted from the spectrum

of the signal. Moreover, in our model we do not use the instantaneous phases but instead the Relative Phase Shifts (RPS's) are used (Saratxaga et al., 2009). The RPS's are the difference between the initial phase shift of every harmonic sinusoid with respect to the first harmonic (F0). They can be calculated from the instantaneous phase of the harmonics using the expression:

$$\theta_k = \varphi_k(t_a) - k\varphi_1(t_a) \tag{9}$$

where θ_k is the RPS, φ_k the instantaneous phase of the k-th harmonic, φ_1 the instantaneous phase of the fundamental frequency harmonic and t_a the instant chosen for the analysis. The result of this formula is wrapped to values inside the $[-\pi, \pi]$ interval.

The RPS's exhibit some desirable properties for the phase representation. The differences of the initial phase shifts of the sinusoidal components determine the actual waveform shape of the signal. Therefore, the RPS's are constant while the waveform shape keeps stable. Furthermore, the RPS's reveal a structured pattern in the phase information of the voiced segments, which is not clear at all in the instantaneous phase representation as it is depicted in fig. 1.

Fig. 1 shows the different phase information for a voiced signal containing five vowels /aeiou/ (fig. 1.c). Fig. 1.a shows the evolution of the usual instantaneous phase both in frequency (vertical axis) and in time (horizontal one), where no structure can be appreciated. Fig 1.b shows the evolution of the RPS's representation for the same signal, where the subjacent phase structure of every vowel is exposed.

As mentioned before, the instantaneous phases are taken from the phases of the windowed signal spectrum at the harmonic frequencies. The spectrum is calculated for every frame by means of an FFT. Afterwards, the instantaneous phases at the frequencies of every harmonic are taken and their phase difference with respect to F0 is computed applying expression (9). For the F0 itself, its instantaneous phase is kept in order to allow a synchronous reconstruction of the original signal.

2.1.3. Voiced/unvoiced band decision

Till this point, we have two independent and complete models of the signal spectrum, one harmonic and the other noise-like. The final stage of the analysis involves deciding whether each band should be represented by the harmonic or by the noise component. The band modelling error is used as input for the decision. As stated in (Griffin & Lim, 1988) the error expression (5) is biased towards the longer periods, for the longer the period is, the more densely the spectrum is sampled, consequently reducing the value of the error. An unbiased expression of the error, proposed in the same paper, is used:

$$\varepsilon_{UB} = \frac{\int_{a_k}^{b_k} \left| S_w(\omega) - \hat{S}_w(\omega) \right|^2 d\omega}{\left(1 - P\sum_n w^4[n]\right) \int_{a_k}^{b_k} \left| S_w(\omega) \right|^2 d\omega} \tag{10}$$

where P is the period of the pitch and $w[n]$ are the samples of the window.

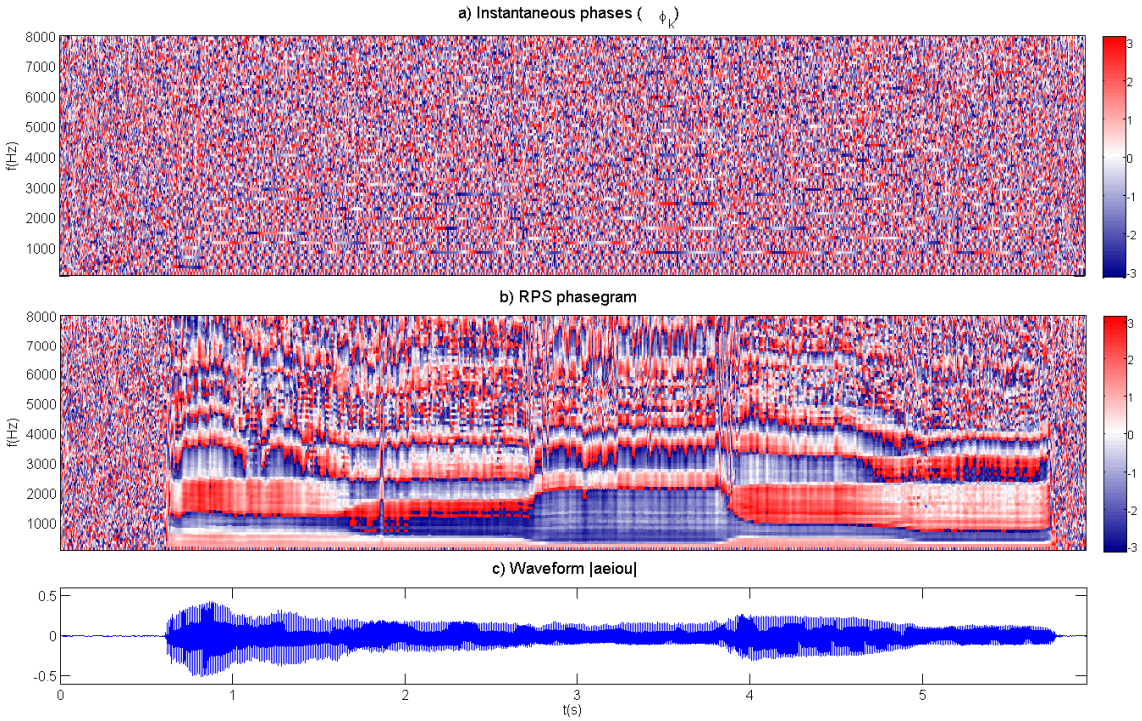

Figure 1. Instantaneous phase vs. RPS phasegrams

This expression gives a normalized error independent of the pitch and of the actual energy of the frame. Expression (10) is calculated using both harmonic and noise models for $\hat{S}_w$ and the band is classified as voiced or unvoiced by comparing the errors produced by each model. A weight can also be used to bias the decision towards one or the other model. In our implementation, the voiced decision (i.e. harmonic component) has been favoured, because it produces perceptually clearer resynthesis.

2.2 HNM-MBE synthesis

The synthesis from the data obtained in the MBE analysis is carried out in two independent processes for the harmonic and the noise components. Both of them are added at the end of the frame generation process.

2.2.1. Synthesis of the harmonic component

The synthesis of the harmonic part requires the pitch, the harmonic coefficients, the V/UV band decision, the phase differences and the instantaneous phase of F0 to be accomplished.

Each frame is synthesized taking into account the initial parameters (*i*) and the final ones, which correspond to the next analysis frame (*i+1*) to ensure continuity. Between these parameters, linear interpolation is used to obtain the amplitudes, RPS's and frequencies for every sample. When a band is voiced (i.e. modelled by a harmonic sinusoid) at the beginning of the frame and becomes unvoiced at the end, or vice versa, the final (or initial) amplitude is set to zero so that the harmonic component fades (or appears) smoothly. As the final parameters will become the initial ones of the next synthesis frame, continuity is ensured. The expression of the harmonic component for frame *i* is:

$$h^i[n] = \sum_{k=1}^{N} A_k[n]\cos\left(\varphi_k[n]\right) \tag{10}$$

where $h^i[n]$ represents the harmonic part of the i-th frame, and N stands for the number of bands of the frame (the greater of the initial and final number of bands). $A_k[n]$ is the linearly interpolated amplitude for each band (k) from its value in the *i*-th frame to its value in the *i+1*-th one. $\varphi_k\,[n]$ is the instantaneous phase and it is function of the time-varying frequency and RPS's.

$$\varphi_k[n] = 2\pi nkf[n] + \theta_k[n] \tag{11}$$

$\varphi_k\,[n]$ is calculated by linearly interpolating both frequency ($f[n]$) and the RPS's ($\theta_k[n]$). The procedure is thoroughly explained in (Saratxaga et al., 2009).

2.2.2. Synthesis of the noise component

The noise component is synthesized by means of a FFT filter. A synthetic spectrum for the white noise is generated first, long enough to minimize the windowing distortion. This length is variable and depends on the required frequency resolution (that is to say, the number of bands) and on the length of the signal to be generated.

Interframe discontinuities are seldom perceptible in noisy signals. Thus, a simple average is done between the noise coefficients of the initial and final analysis frames and they are kept constant within the frame. In an analogous but inverse way to the harmonic part, if a band starts as unvoiced but ends as voiced (or on the contrary, becomes unvoiced) the corresponding unvoiced coefficient is set to zero.

The rules above are applied for each band, and a spectral envelope is obtained. Then, it is applied to the synthetic noise spectrum and the inverse Fourier transform is calculated.

$$n[n] = \mathcal{F}^{-1}\left\{E(\omega)\cdot W(\omega)\right\} \tag{12}$$

where $E(\omega)$ is the envelope and $W(\omega)$ the synthetic spectrum.

The last step of the synthesis process is the addition of the harmonic and noise signals to get the complete frame, which is concatenated to the previously generated output signal.

2.3 Pitch and duration modifications

Changing most of the parameters of the model (amplitudes, phases or banding decisions) has an immediate impact on the signal spectrum. On the contrary changing the pitch or the duration of the signal should ideally leave the spectrum unaffected, but imply a different kind of parameter modification.

Duration changes using RPS's are immediate. They just imply changing the number of synthesis samples per analysis frame according to the length modification factor, while the rest of the parameters remain unaffected.

By the contrary, pitch changes have deeper effects, because modifying the pitch implies modifying the frequencies of all the harmonic components and thus the number of parameters. The problem is how to estimate the values of the parameters at the new frequencies of the harmonics departing from the original ones. The usual solution (Quatieri & McAulay, 1992) consists in considering the original parameters as points of a frequency envelope which is resampled at the new frequencies to obtain the new set of parameters. AhoTransf uses linear interpolation to obtain the new parameters and employs this technique both for amplitudes and for the RPS's.

3. AhoTransf

AhoTransf is a modular tool designed to visualize and modify the parameters of harmonic speech models. It also integrates speech analysis and resynthesis along with the GUI, thus allowing a straightforward manipulation of the speech signal. The tool has been developed using the HNM-MBE model, but other harmonic models can be integrated with little effort.

The application is developed in Matlab and is organized around three core modules: the director module, the displaying module and the editing module. Around them the HNM-MBE analysis, synthesis and modification algorithms' implementations are used to get and process the data. This modular structure allows using the tools core modules not only with HNM-MBE parameters but also different harmonic models with minimal modifications. A diagram of the structure of the tool is

shown in fig. 2.

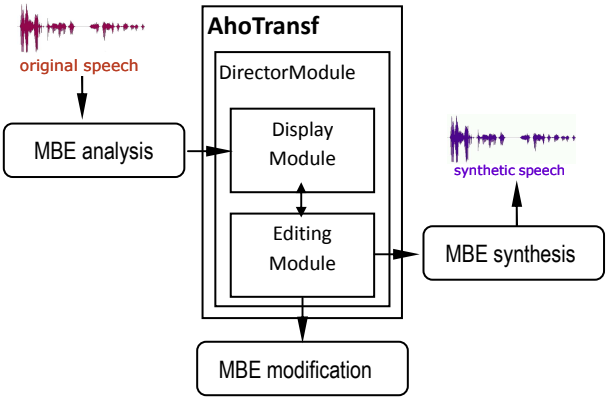

Figure 2. Modular structure of AhoTransf

The director module captures the user commands and manages the invocation of the rest of the components and functions in order to fulfil them. We will now describe the functionalities of the display and editing modules as they gather the main functionalities of the tool.

3.1 Display Module

The display module is responsible of formatting the parameters of the model so that they can be easily interpreted by the user. The display module has a parametric and modular structure that allows an effortless reconfiguration to adapt it to other kind of speech models.

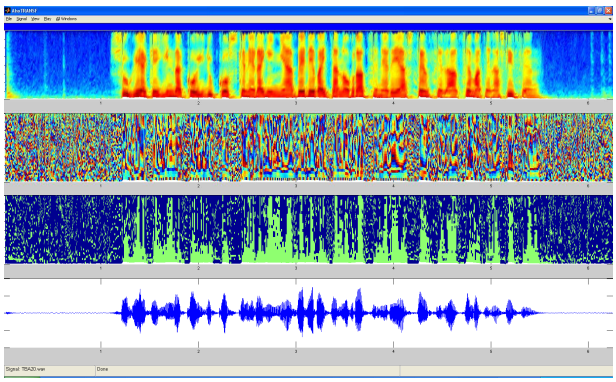

Figure 3. Visualization window

For the HNM-MBE model, the visualization window shows four panels. Three of the panels show representations of the amplitudes and phases of the harmonic part of the model, and the frame-by-frame voicing decision. The fourth one shows the signal waveform. In order to keep the window as simple as possible, the amplitudes of the noise part are not shown because they look very much like the amplitudes of the harmonic part since they model the same PSD.
The parameters of the model are displayed in spectrogram-like graphics, with time in the horizontal axis and frequency in the vertical one. This representation is not directly obtained from the parameters of the model. In fact, for every analysis frame the number of harmonic parameters is different as they are function of the pitch at the time of the analysis. So the parameters have to be scaled in frequency in order to get a meaningful representation.
The display module provides several visualization facilities such as time axis scrolling, selection and zooming (either individually by panel or combined for all the panels).
Regarding the synthesis possibilities, the user can choose to hear the whole of the signal or parts of it. He or she can also hear the original and the resynthesized signals. In this last case, the user can choose to hear separately the signals corresponding to the harmonic or the noise parts of the model.

3.2 Edition Module

The modification of the parameters to obtain different voice perceptual qualities is a complicated task, as they require non-uniform coordinated modification of whole groups of parameters. The edition module of AhoTransf allows simple but detailed modification of the amplitudes, phases, voiced/unvoiced decisions per band, pitch and overall duration of the signal.
The edition window shows four panels with the harmonic amplitudes, phases, voiced/unvoiced decisions and the pitch of the signal. Modifications can be applied either to the whole signal or to a selected segment. For bidimensional parameters (i.e. those dependent on time and frequency) it is possible to limit the modification to a certain segment and certain frequencies. The selection of the parameters to be changed is easy, as it is done using the mouse. Zooming and hearing tools are available to help with the selection of the desired fragment.
The editing possibilities are different depending on the parameter.

- Amplitudes (A_k, B_k): Changes in this panel are applied to the amplitudes of both the harmonic and noise models. The amplitudes can be set to a certain value and can also be scaled by a frequency dependent factor thus allowing modifying the tilt of the spectrum.
- Phases: They can be adjusted to a frequency dependent mathematical expression to test the perceptual influence of different phase structures. They can also be set to random values.
- Voiced/Unvoiced decisions can be set per band. This feature allows producing pure harmonic or noisy versions of the original signal, and also studying the actual contribution of each component to the voice quality (harmonic-to-noise ratio, breathiness, maximum voicing frequency).
- Pitch can be scaled, interpolated between two certain values or set to a certain value, thus allowing prosodic modifications.
- Signal duration can be scaled by a factor.

Changes in the parameters are immediately visualized at the corresponding panel and finally the signal can be resynthesized using the modified data.

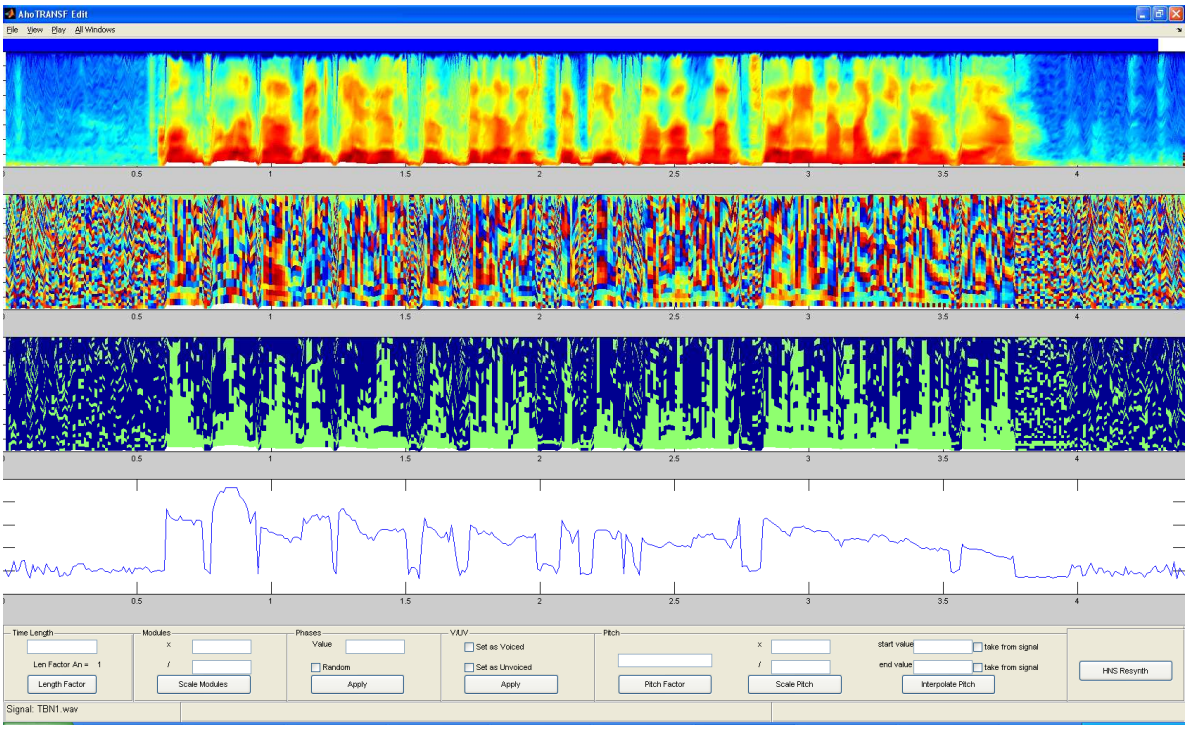

Figure 4. Edition window

4. Conclusions

We have developed a graphic application for the visualization and edition of the parameters of our HNM-MBE model. AhoTransf is a modular application that can be customized to manage different sets of parameters, so it will be expanded to work with other models of the harmonic family.

This application will be used for research purposes to test the perceptual effects of the changes in different parameters, as the GUI provides a quick and effortless way to check them. It will also be used for educational purposes to help explaining the harmonic models.

Future work could be done to include different speech coding algorithms in order to compare their parameters and resynthesis quality.

5. Acknowledgements

The work presented in this paper has been partially funded by the Spanish Government under grant TEC2009-14094-C04-02 (BUCEADOR project) and by the Basque Government under grant IE09-262 (BERBATEK project).

6. References

Dutoit, T., Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication* 13, (3-4), pp. 435--440.

Erro, D., Moreno, A., Bonafonte, A. (2007). Flexible harmonic/stochastic speech synthesis. *Proceedings of the 6th SSW6*. Bonn, Germany.

Griffin, D.W., Lim, J. (1988). Multiband Excitation Vocoder. *IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-36*, 36(8), pp.1223--1235.

Laroche, J., Stylianou, Y., Moulines, E. (1993). HNM: a simple, efficient harmonic+noise model for speech. *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 169--172.

Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., Sainz, I. (2007). Evaluation of pitch detection algorithms under real conditions. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, 4, pp. 1057--1060.

Quatieri, T., McAulay, R. (1986). Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(6), pp.1449-–1464.

Quatieri, T., McAulay. R. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing* 40(3), pp. 497--510.

Saratxaga, I., Hernáez, I., Erro, D., Navas, E., Sánchez, J. (2009). Simple representation of signal phase for harmonic speech models. *Electronics Letters* 45(7), pp. 381-383.

Stylianou, Y. (1996). Harmonic plus Noise models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD Thesis. Ecole Nationale Supérieure des Télécommunications. Paris.