MODELING SPECTRAL CHANGES IN SINGING VOICE FOR PITCH MODIFICATION

Isabel Barbancho **José L. Santacruz** **Lorenzo J. Tardón** **Ana M. Barbancho**

Universidad de Málaga, ATIC Research Group, Andalucía Tech, E.T.S.I. Telecomunicación,
Dpto. Ingeniería de Comunicaciones, Campus Universitario de Teatinos s/n,
29071, Málaga, Spain

ibp@ic.uma.es, jls@ic.uma.es, lorenzo@ic.uma.es, abp@ic.uma.es

ABSTRACT

We present an advanced method to achieve natural modifications when applying a pitch shifting process to singing voice by modifying the spectral envelope of the audio excerpt. To this end, an all-pole spectral envelope model has been selected to describe the global variations of the spectral envelope with the changes of the pitch. We performed a pitch shifting process of some sustained vowels with the envelope processing and without it, and compared both by means of a survey open to volunteers in our website.

1. INTRODUCTION

Since some years ago there is an increasing interest in voice synthesis systems for entertainment purposes as well as the use of pitch shifting algorithms in music production [1] with creative aims or to correct singer mistakes during a recording process. To this end, the most commonly used algorithms are focused on the spectral envelope preservation in order to achieve a natural transformation modifying as little as possible the original timbre. Nevertheless, it has been proved that the spectral envelope changes as intensity [5] or pitch [3] varies, as in the cases of age or sex. Therefore, spectral envelope preservation is not the best approach for performing wide variations in pitch maintaining the voice quality, thus we decided to study this particular effect with the goal of obtaining more realistic or natural results.

2. SPECTRAL ENVELOPE AND VARIATION ALONG PITCH

The spectral envelope is a smooth function passing through the prominent peaks of the spectrum, and it is generally considered as one of the determining factors for the timbre of a sound. Among the many existing approaches to parameterize the spectral envelope, most of them are based on Linear Prediction Coding (LPC), we opted for a formant-based model developed in a previous work [6]. This spectral envelope model is inspired by some speech/singing synthesis systems like [2], and it is based on considering several resonator filters in parallel with a certain envelope slope, for modeling the acoustic formants and glottal source, respectively.

We modeled the spectrum of a sung vowel in the frequency band [0, 5000 Hz] with a source filter and a set of five parallel resonators, which are the glottal resonator and the first four formants of the vocal tract.

The following table shows the ranges of values for the different parameters in the model:

Parameter	Range	Parameter	Range
$Gain_{dB}$	[-200, 0] dB	F_2	[500, 3000] Hz
$SlopeDepth_{dB}$	[-50, 100] dB	B_2	[40, 1000] Hz
F_{GP}	[0, 600] Hz	F_3	[500, 3000] Hz
B_{GP}	[100, 2000] Hz	B_3	[40, 1000] Hz
F_1	[150, 1100] Hz	F_4	[3000, 5000] Hz
B_1	[40, 1000] Hz	B_4	[100. 1000] Hz

Table 1. Range of values used in the proposed spectral envelope model.

where F_i and B_i are the the central frequency and the 6 dB-bandwidth of resonator i, respectively. Likewise, F_{GP} and B_{GP} refer to the center frequency and the and 6 dB-bandwidth of the glottal pulse.

Through a dataset consisting of 76 sustained notes sung by two male and two female amateur singers, we conducted an analysis of the variation of the parameters presented before with the pitch using a linear least-squares regression. All these variables have been mixed, in order to find a global tendency of the variation. Among all the parameters for our model of the spectral envelope, we have found significant evidence of the variation with the pitch especially in the center frequency of first formant (F1). As shown in Fig. 1, F1 increases with the pitch (F0). Taking into account this tendency, a linear approximation has been found by using robust regression, which is more accurate than ordinary least squares in this particular case. Thus, some outliers are rejected and the resulting slope is more reliable. The straight line shown in the figure is modeled by the following expression:

$$F_1' = F_1 + \Delta F0 \cdot \omega \quad (1)$$

where F_1' is the new value of center frequency of the first formant, F_1 is the original value, and ω is the variation weight or slope of the linear transformation defined. This weight has similar but different values for the harmonic and the residual component of the signal, as it is shown in Table 2.

parameter	ω (harm.)	ω (res.)
F1	0.89	0.84

Table 2. Slope of the linear transformation with pitch shifting of F1 with respect to the frequency displacement for both the harmonic and residual audio components.

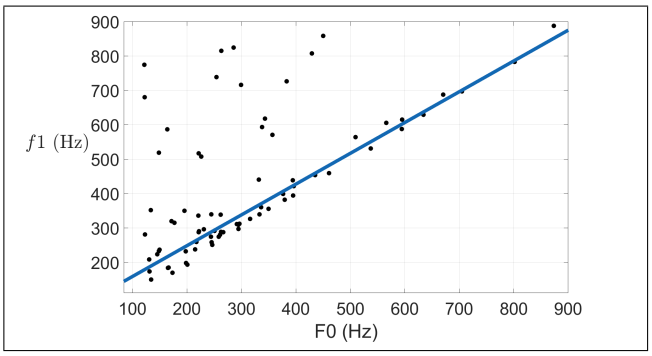

Figure 1. F1 evolution along pitch, harmonic component.

3. PITCH SHIFTING ALGORITHM

Some parameters must be extracted in order to perform the voice transformation as described, and these are the spectral envelope and the F0 of the signal. Pitch detection is accomplished by using the YIN algorithm [4] to extract the F0 at frame-level. The F0 is also necessary to determine the position of the $pitch\ marks$, which provide the centers of the segmentation windows in PSOLA algorithm [7]. The spectral envelope $H_1(z)$ of the original sound is estimated by making use of the parametric model presented.

With these parameters extracted, the pitch is increased or decreased by using the PSOLA technique with a desired skip or leap. The time domain pitch synchronous overlap-add technique (PSOLA) is one of the most used methods due to its quality and simplicity.

Once we have generated a pitch shifted version of the original signal, which has been obtained by using a formant preservation technique, it is necessary to modify the spectral envelope in order to achieve a more natural result. To achieve this, the first formant, F1, in the original spectral envelope $H_1(z)$ is shifted according to Eqn (1), to produce a new transformed spectral envelope $H_2(z)$. Then, the signal is re-synthesized using the new spectral envelope to give the final transformed signal. According to Table 2, this process is carried out for both the harmonic and residual components separately.

4. EVALUATION AND RESULTS

The system performance was subjectively evaluated by means of a survey open to volunteers in our website [1].

[1] http://www.atic.uma.es/PitchShifting/PitchShiftingSurvey/

The 18 participants blindly listened to 6 different transformations in random order and they were asked to evaluate the quality of each one, with regard to the original audios. These synthesized sounds were open vowels (/a/) and closed vowels (/i/). As Fig. 2 shows, the majority of evaluators considered that the quality of our method was better than the simple PSOLA transformation for every single sample. In general, our approach seems to achieve better results for closed vowels (/i/) probably because of the larger separation between formants 1 and 2. Moreover, higher notes obtained better results due to the proportionality between the shifting and the frequency. Both cases make the transformation distinguishable from an ordinary pitch-shifting method.

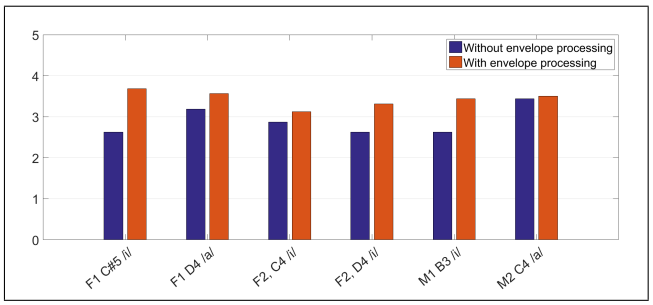

Figure 2. Results of the survey.

The main effects that achieved a better result were the higher vowel intelligibility and the lower breathness of our method. With the ordinary PSOLA, some of the notes lost the vowel intelligibility and sound like a neutral vowel. Also, with the simple PSOLA, a small frequency delay appears between the harmonic and the residual component at the first formant, producing the breathness especially in higher notes.

5. CONCLUSIONS

The system developed attempts to overcome the timbre variations problem along pitch, due to most of the system developed until now being based in formant preservation when pitch-shifting. A first analysis allowed us to define a function describing the variation of formants along pitch in singing voice. A set of vowels have been pitch-shifted using a PSOLA algorithm and the envelope processing through the obtained function.

A subjective evaluation of the system performance was carried out by means of an open survey, comparing our method with the simple PSOLA. The results showed that our approach is more accurate due to the envelope processing maintaining the vowel intelligibility and reducing the breathness, especially at higher pitch.

6. ACKNOWLEDGEMENTS

This work has been funded by the Ministerio de Economía y Competitividad of the Spanish Government under Project No. TIN2013-47276-C6-2-R and by the Junta de Andalucía under Project No. P11-TIC-7154. The work has been done at Universidad de Málaga. Campus de Excelencia Internacional Andalucía Tech.

7. REFERENCES

[1] Celemony software: Melodyne editor. *http://www.celemony.com*.

[2] Jordi Bonada, Òscar Celma, Àlex Loscos, Jaume Ortolà, Xavier Serra, Yasuo Yoshioka, Hiraku Kayama, Yuji Hisaminato, and Hideki Kenmochi. Singing voice synthesis combining excitation plus resonance and sinusoidal plus residual models. In *Proceedings of International Computer Music Conference*, 2001.

[3] Kateřina Chládková, Paul Boersma, and Vclav Jon Podlipsk. Online formant shifting as a function of f0. In *Proc. of Interspeech 2009*, pages 464–467, 2009.

[4] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[5] J. E. Huber, G. M. Curione Stathopoulos, T. A. Ash, and K. Johnson. Formants of children, women and men: The effects of vocal intensity variation. *The Journal of the Acoustical Society of America*, 106:1532, 1999.

[6] Emilio Molina, Isabel Barbancho, Ana M. Barbancho, and Lorenzo J. Tardón. Parametric model of spectral envelope to synthesize realistic intensity variations in singing voice. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2014)*, pages 634–637, 2014.

[7] Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467, 1990.