International Journal of
Applied Mathematics,
Electronics and Computers

IJAMEC

ISSN: 2147-8228 www.atscience.org/IJAMEC

Advanced Technology and Science

Original Research Paper

Interface for Dynamic Modification of the Transformation Parameters of the PSOLA Algorithm#

Lyes Demri*, Leila Falek, Hocine Teffahi [1]

***Abstract*:** The prosody of a speech signal is related to many factors: the social and geographical origin of the speaker, his or her emotional state, his physiological state (weariness, sickness, …) and the type of the sentence (interrogative, affirmative, etc.). A good synthesis or speech transformation system must account for all of these factors in order to produce a speech that sounds natural. In this paper, we propose a graphical interface for the modification of the prosodic features of the speech signal (the melodic curve - fundamental frequency and temporal
organization of the syllables - and the formantic trajectories) using the PSOLA algorithm. The interface allows the user to manually introduce the desired trajectories of the transformation parameters of the PSOLA algorithm in order to produce a transformed signal which has the desired prosody. The results are acceptable, especially for the modification of the fundamental frequency and of the temporal organization of the source signal.

***Keywords*:** *Speech Synthesis, PSOLA Algorithm, PSOLA Parameters, Prosodic Parameters, Dynamically Changing, Interface*

1. Introduction

The ability to modify the acoustic features of a voice is an important factor in voice transformation and synthesis systems. It is commonly accepted that a large part of the information contained in a speech signal lies in the prosody (the melodic curve, that is, the evolution of the fundamental frequency F0 and the temporal organization of the syllables), and in the voice quality (represented by the formants of the spectral envelope) ([1], [2]). These features inform us both about the type of sentence (affirmative, interrogative, exclamative), and the emotional state of the speaker (sadness, happiness, anger, etc.).
In this study, we focus on commanding those features through an interactive interface between the user and the transformation system which is built on the PSOLA algorithm. The PSOLA algorithm is based on the calculation of three transformation parameters, α, β and γ, which correspond respectively to the temporal organization, the fundamental frequency, and the voice quality.

2. The PSOLA Algorithm

The PSOLA algorithm (Pitch Synchronous Overlap-Add) is well known in the domains of voice transformation and synthesis. It is used to modify the intrinsic characteristics of a voice signal [4], [8]. The PSOLA algorithm is based on an analysis/synthesis scheme and can modify the signal both in the time and the frequency domains [3]. Here is a brief description of the algorithm.

2.1. Analysis

The analysis step consists of placing reading marks mi on the original signal s(t), according to the local characteristics of its components (periodic, random, transitory). These marks are placed on the signal's local maxima of energy, and usually correspond to the impulsions in the source-filter scheme, and to the glottal impulsions in the case of a voice [9]. The signal is thus segmented into elementary signals through the use of Hanning windows hi(t) whose lengths are usually 2 or 4 times that of the fundamental frequency and centered on the mi [10]. Each of the windows must be centered on the the local maximum of energy. The resulting signal si(t) can be written as :

$$s_i(t) = h_i(t - m_i)s(t) \quad (1)$$

The analysis algorithm is as follows (figure 1):
- Determination of the pitch period P(t) of the input signal and of time instants (pitch marks) ti. These pitch marks are in correspondence with the maximum amplitude or glottal pulses at a pitch synchronous rate during the periodic part of the sound and at a constant rate during the unvoiced portions. In practice P(t) is considered constant P(t) = P(ti) = ti+l - ti on the time interval (ti, ti+l).
- Extraction of a segment centered at every pitch mark ti by using a Hanning window with the length Li = 2P(ti) (two pitch periods) to ensure fade-in and fade-out.

2.2. Synthesis

The synthesis step consists in placing writing marks which maintain the same distances between each other as those of the reading marks, so that the original fundamental frequency is preserved. The correspondence index, which is the corresponding position to the writing marks in the original signal, allows us to select the elementary signal which will be used at this writing mark (an already existing elementary signal (TD-PSOLA), or interpolated in the time domain (TDI-PSOLA), or in the frequency domain (FDI-PSOLA) [10]).
Synthesis algorithm (figure 2): for every synthesis pitch mark $\widetilde{tk}$
1. Choice of the corresponding analysis segment i (identified by the time mark ti) minimizing the time distance $|\alpha ti - \widetilde{tk}|$.
2. Overlap and add the selected segment. Notice that some input segments will be repeated for α > 1 (time expansion) or discarded when α < 1 (time compression).

[1] *Electronics and Computer Science Department, USTHB,BP 32, Bab-Ezzouar, Algiers, Algeria.*
** Corresponding Author: Email: ldemri1987@hotmail.fr*
 # This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

3. Determination of the time instant $\widetilde{tk+1}$ where the next synthesis segment will be centered, in order to preserve the local pitch, by the relation

$$t_{k+1} = \tilde{t}_k + \tilde{P}(\tilde{t}_k) = \tilde{t}_k + P(t_i) \quad (2)$$

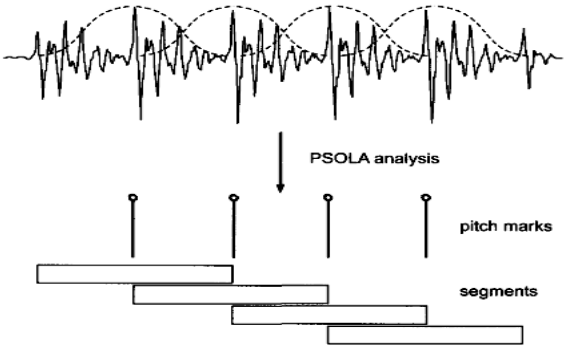

Figure 1. PSOLA pitch analysis

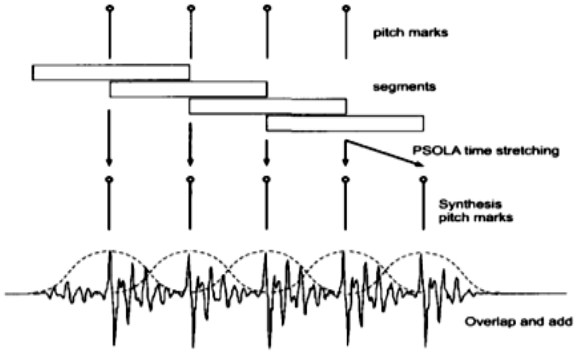

Figure 2. PSOLA synthesis for time stretching

It should be noticed that the determination of the pitch and of the position of pitch marks is not a trivial problem and could be difficult to implement robustly in real-time. Stretching factors typically range from α=0.25 to 2 for speech. Audible buzziness appears in unvoiced sound when larger values are applied, due to the regular repetition of identical input segments. In order to prevent the algorithm from introducing such an artificial short-term correlation in the synthesis signal, it is advisable to reverse the time axis of every repeated version of an unvoiced segment. With such an artifice, speech can be slowed down by a factor of four, even though some tonal effect is encountered in voiced fricatives, which combine voiced and unvoiced frequency regions and thus cannot reversed in time.

A further effect that can be obtained by a variation of PSOLA is linear scaling of formant frequencies (see figure 3). In fact, we saw that a time scale of a signal corresponds to an inverse frequency scale. Thus when we perform time scaling of the impulse response of a filter, we inversely scale the frequency of formants. In PSOLA terms, this corresponds to time scaling the selected input segments before overlap and add in the synthesis step, without any change in the pitch marks calculation. To increase the frequencies of formants by a factor γ, every segment should be shortened by a factor 1/γ by resampling. For example, the average formant frequencies of female adults are about 16 percent higher than those of male adults, and children's formants are about 20 percent higher than female formants. Notice that care should be taken when the frequencies increase in order to avoid foldover. Ideally band-limited resampling should be used [12].

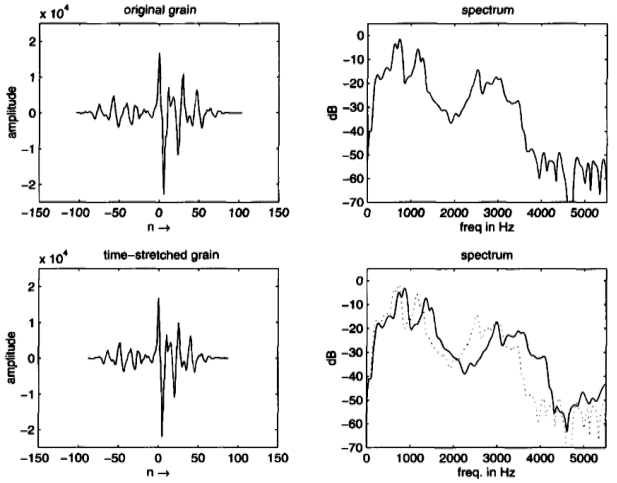

Figure 3. PSOLA: Variation of PSOLA as linear formant scaling

3. Presentation of the Interface

We have realized an interactive interface between the user and the computer that allows a dynamic modification of the transformation parameters of the PSOLA algorithm. The objective of the interface is to obtain a transformed signal from a source signal, through the modification of the three parameters cited above. The interface's inputs are:

the source signal

the trajectories for the three transformation parameters α, β and γ.

The interface outputs the transformed signal. Here are the steps required to produce the transformed signal.

3.1. Dynamic Modification of the Fundamental Frequency

During the reconstruction step, the algorithm must calculate where the next waveform is to be placed. The position is given by the relation:

$$t_k = t_k + (T/\beta) \quad (3)$$

Where:

t_k is the position of the local maximum of the current waveform,

T is the period of the source signal at this instant,

β is the shifting factor of the pitch.

For β = 1, we therefore have:

$$t_k = t_k + T \quad (4)$$

In order to change the pitch in a dynamic way, we must introduce β as varying during the reconstruction phase :

$$t_k = t_k + (T/\beta(i)) \quad (5)$$

where β is now a vector whose values are obtained by a simple linear interpolation between an initial point (β_ini) and a final point (β_end). It is then necessary to send in only those two values to the PSOLA function which will automatically compute the values β(i) of β between these 2 points. With these values, the PSOLA algorithm can realize a dynamic modification of the pitch.

3.2. Modification of α

It is not necessary for α to change dynamically like β. To illustrate this, let us take the example of a speech signal that contains 4 syllables « taratata » (/ta/, /ra/, /ta/, /ta/) that have the same lengths

(d1 = d2 = d3 = d4). If we apply a dynamic α that varies linearly from 1 to 2, the 4 syllables will have durations such that d1 < d2 < d3 < d4. The interface will behave just as if it segmented the original signal into 4 segments and applied a different value of α for each segment. It is therefore not necessary to introduce a vector of α(i) but simply to segment the input signal into many parts and apply a different α for each segment.

3.3. Dynamic Modification of the Voice Quality (γ)

The dynamic modification of the voice quality is similar to that of the pitch (β). The user sends in a couple of values γ_ini and γ_end, and a linear interpolation between these two points generates the vector γ that contains the values γ(i) for each waveform.

4. Experimental Results

The interface has been realized in MATLAB and is illustrated in figure 4.

4.1. Functioning Principle of the Graphical Interface

The interface shows 4 graphs:
The speech signal
The desired variation curve for α
The desired variation curve for β
The desired variation curve for γ
The desired trajectories for α, β and γ can be realized either by manually introducing the values facing each graph, or by directly clicking on the graphs using the mouse where the points should be. Once the trajectories are introduced, the user can click « play » which will transform the source signal and playback the result. The interface first applies β, then γ, and then α (in this order). For each of the parameters, the signal is decomposed into segments. On each segment, a linear transformation is applied, as described in section 3. Once all the segments are transformed, they are concatenated to obtain the transformed signal. For each parameter the interface allows to save, load or erase the values of the parameters. Other click buttons allow the user to :
Select file: select a source signal in a directory.
Save transformed signal : save the transformed signal
Play : Transform the source signal and play the result
Play original signal : plays the source signal.

4.2. Results Illustrations

In this part, we present a few results obtained using the interface. The aim was to try and reproduce the prosody of a target sentence on a sentence that has equal length and same pitch syllables.
The target sentence is in french « oui pourquoi pas ?» uttered on the interrogative tone. The sentence is first recorded with the interrogative prosody - this will be the target sentence. The sentence is then recorded again on a neutral and monotonic tone (constant fundamental frequency and speech speed), as our source signal. The objective is to imitate the target by transforming the source file. The results obtained for the pitch and the temporal organization are given in figure 4.

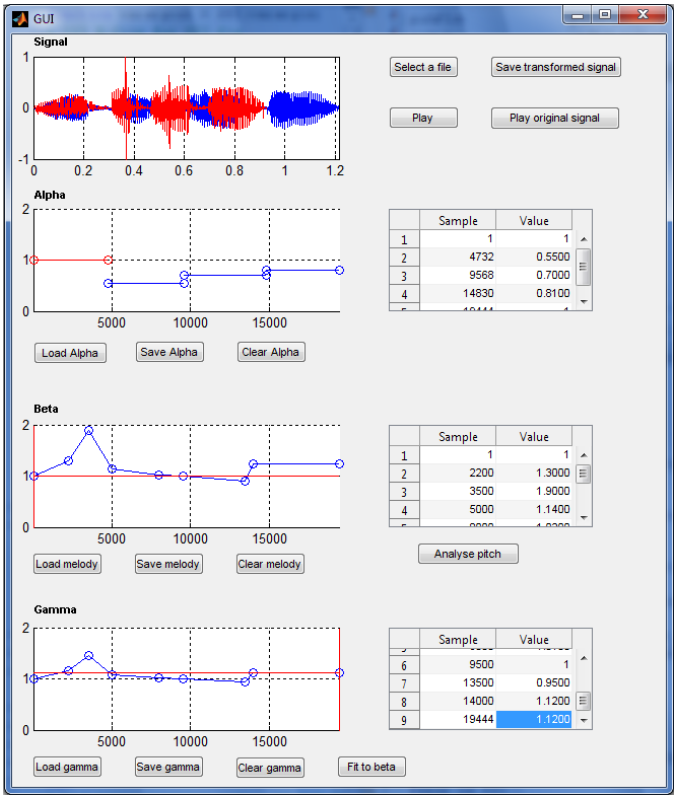

Figure 4. Matlab interface. Transformation for neutral to interrogative tone (red : source signal, blue : transformed signal)

Figure 5 and Table I show that:
- The syllables of the source signal were shortened, and have approximately the same durations as those of the target signal.
- The pitch of the transformed signal is close to that of the target signal.

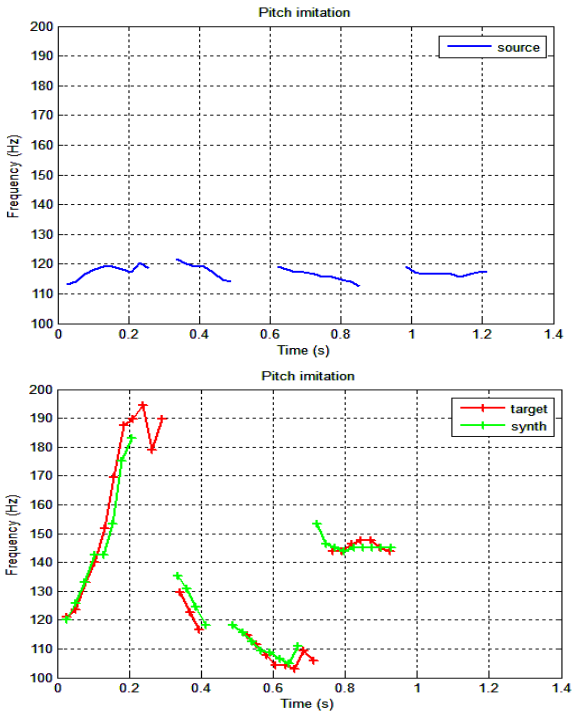

Figure 5. Melodic curves obtained (blue : source signal, red : target, green : transformed signal)

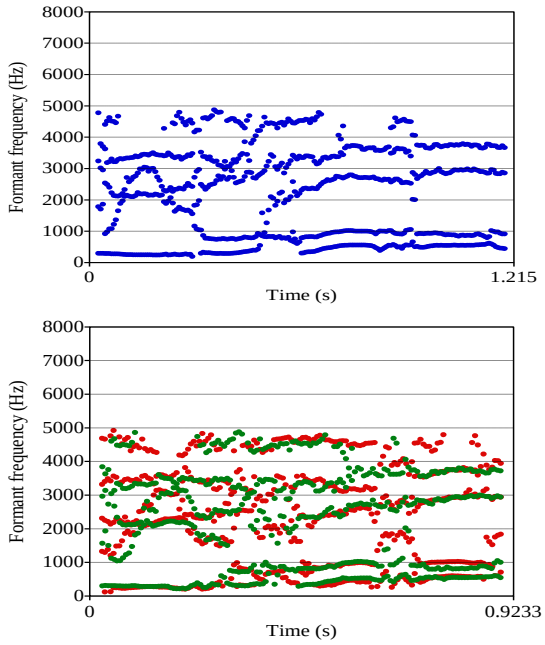

Figure 6. Formantic trajectories (blue: source signal, red: target signal, green: transformed signal)

The pitch was calculated via the autocorrelation method:

$$\hat{R}_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n^* & m \geq 0 \\ \hat{R}_{yx}(-m) & m < 0 \end{cases} \quad (6)$$

with $x = y$. This method is based on the fact that the autocorrelation of a periodic signal is itself periodic, with the same period. With this knowledge, it is possible to determine the pitch of a vocal signal by looking for the maximum of the autocorrelation function of the vocal signal with itself.

Figure 6 and table II show that the formants of the transformed signal are also close to those of the target signal.

The formants are obtained using PRAAT®. The algorithm for computing the formants is to resample the sound to a sampling frequency of twice the value of *Maximum formant* (which is an adjustable setting). Pre-emphasis is then applied, and then for each analysis window, Praat applies a Gaussian-like window, and computes the LPC coefficients with the algorithm by Burg, as given by [13] and [14] Press et al. (1992).

These results are similar to the ones reported in [11], in which the PSOLA algorithm is used to convert a female voice into a male voice. In this study however, the modification factors of the PSOLA parameters are not constant in time.

Table 1. Fundamental frequencies for the 3 Signals

Target		Source		Synth	
t (s)	f0(Hz)	t (s)	f0(Hz)	t (s)	f0(Hz)
0.0264	120	0,0258	113	0,0262	118
0.0528	123	0,0517	114	0,0524	126
0.0791	132	0,0775	116	0,0786	134
0.1055	139	0,1034	118	0,1048	141
0.1319	151	0,1292	119	0,1310	145
0.1583	169	0,1551	119	0,1573	151
0.1847	187	0,1809	118	0,1835	177
0.2110	189	0,2068	117	0,2097	187
0.2374	194	0,2327	120	0,2359	NaN
0.2638	179	0,2585	118	0,2621	NaN
0.2902	189	0,2844	NaN	0,2883	NaN
0.3166	NaN	0,3102	NaN	0,3146	NaN
0.3429	129	0,3361	121	0,3408	132
0.3693	122	0,3619	120	0,3670	127
0.3957	116	0,3878	119	0,3932	122
0.4221	NaN	0,4137	119	0,4194	120
0.4485	NaN	0,4395	117	0,445	NaN
0.4748	300	0,4654	114	0,4719	NaN
0.5012	NaN	0,4912	114	0,4981	NaN
0.5276	114	0,5171	NaN	0,5243	115
0.5540	111	0,5429	84	0,5505	112
0.5804	107	0,5688	NaN	0,5767	110
0.6067	104	0,5946	NaN	0,6029	107
0.6331	104	0,6205	119	0,6292	105
0.6595	103	0,6464	118	0,6554	105
0.6859	109	0,6722	117	0,6816	NaN
0.7123	105	0,6981	117	0,7078	NaN
0.7387	NaN	0,7239	116	0,7340	145
0.7650	143	0,7498	115	0,7602	144
0.7914	143	0,7756	115	0,7865	143
0.8178	146	0,8015	114	0,8127	143
0.8442	147	0,8274	114	0,8389	143
0.8706	147	0,8532	112	0,8651	143
0.8969	145	0,8791	NaN	0,8913	145
0.9233	143	0,9049	NaN		
		0,9308	NaN		
		0,9566	NaN		
		0,9825	119		
		1,008	117		
		1,034	116		
		1,060	116		
		1,085	116		
		1,111	116		
		1,137	115		
		1,163	116		
		1,189	117		
		1,215	117		

Table 2. Formants of the synthetic signal

Target					Source					Synth				
time (s)	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	time (s)	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	time (s)	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
0,111	263	2121	2625	3495	0,107	254	2085	2545	3332	0,108	285	2090	2335	3311
0,136	270	2190	2971	3373	0,132	246	2148	2791	3385	0,133	279	2130	2595	3352
0,161	258	2271	3092	3567	0,157	244	2179	2968	3435	0,158	266	2190	3082	3456
0,186	236	2299	3018	3553	0,182	242	2164	2877	3449	0,183	229	2196	3108	3472
0,211	218	2302	2426	3308	0,207	245	2148	2571	3428	0,208	259	2150	2942	3448
0,236	230	2004	2377	3452	0,232	251	2091	2297	3414	0,233	303	2033	2464	3439
0,261	231	1760	2352	3372	0,257	254	1725	2342	3366	0,258	361	1558	2345	3409
0,286	284	1703	2779	3285	0,282	243	1421	2366	3413	0,283	385	1490	2486	3461
0,311	706	2404	3134	3773	0,307	298	1232	2490	4264	0,308	356	983	2516	3397
0,336	267	771	2607	3488	0,332	324	869	2463	3464	0,333	343	829	2335	3126
0,361	266	705	2678	3156	0,357	294	754	2559	3203	0,358	362	957	2287	3407
0,386	641	2016	2972	4574	0,382	289	754	2750	3091	0,383	423	951	2321	3474
0,411	722	2046	3432	4665	0,407	297	756	2731	3151	0,408	333	865	2409	3352
0,436	698	2082	3419	4680	0,432	323	772	2677	3342	0,433	329	838	2365	3370
0,461	714	759	2201	3457	0,457	358	810	2864	3720	0,458	340	818	2355	3317
0,486	343	820	2293	3457	0,482	428	958	2888	4331	0,483	385	824	2395	3230

0,511	314	866	2403	3362	0,507	834	1599	3056	4428	0,508	464	887	2487	3349
0,536	357	907	2471	3157	0,532	833	1942	3194	4442	0,533	510	979	2549	3413
0,561	418	957	2541	3147	0,557	760	2126	3218	4468	0,558	526	1014	2705	3634
0,586	416	983	2594	3174	0,582	669	2205	3374	4461	0,583	519	1006	2732	3649
0,611	394	980	2581	3199	0,607	336	816	2412	3433	0,608	489	975	2705	3645
0,636	937	2272	3139	4551	0,632	349	825	2378	3311	0,633	486	986	2640	3646
0,661	618	1821	2816	4009	0,657	414	858	2456	3275	0,658	499	920	2677	3701
0,686	794	1920	3175	3778	0,682	470	927	2618	3365	0,683	509	898	2788	3651
0,711	482	986	2759	3571	0,707	510	978	2714	3421	0,708	538	869	2813	3616
0,736	523	995	2836	3606	0,732	547	1008	2750	3650	0,733	541	815	2832	3678
0,761	563	1008	2873	3674	0,757	556	1010	2750	3717	0,758	543	818	2885	3743
0,786	583	1016	2910	3731	0,782	555	1001	2720	3676	0,783	549	816	2940	3759
0,811	583	1016	2936	3767	0,807	538	986	2714	3647					
					0,832	494	969	2700	3625					
					0,857	487	991	2643	3616					
					0,882	534	958	2622	3610					
					0,907	578	949	2558	3658					
					0,932	503	949	2650	3715					
					0,957	502	925	2792	3695					
					0,982	532	928	2884	3703					
					1,007	547	916	2896	3728					
					1,032	548	886	2885	3740					
					1,057	551	865	2945	3761					
					1,082	551	865	2961	3760					
					1,107	559	881	2948	3720					

5. Conclusion

The results obtained with the PSOLA algorithm and the interface are pretty acceptable. The interface has been used to produce various types of sentences. The original sentences were recorded with constant pitch and speech speed, and the modifications of pitch, voice quality and temporal organization were then applied to the signal using the interface.

The voice quality, however, is the most difficult feature to modify using this method. A mere resampling of the waveforms only allows a basic modification of the spectral envelope. It is not possible to alter one formant independently of the others.

Future works are focused on optimizing the interface so that it will be able to automatically imitate the prosodic features of a target signal.

The advantage of this interface lies in its simplicity. It can be used in various domains such as music, cinema, emotion synthesis, correcting speech pathologies, etc.

References

[1] M. Schübiger: *English intonation*. Max Niemeyer Verlag, Tübingen (1958)

[2] D. Crystal: *Prosodic systems and intonation in English*. Cambridge University Press, London (1969)

[3] G. Peeters : *Modèles et modification du signal sonore adaptés à ses caractéristiques locales*, Thesis, Paris (2001)

[4] [4]. H. Valbret, E. Mouline, J. Tubach : *Voice transformation using PSOLA technique*, Speech Communication 11 (1992), p. 175-187.

[5] M. Kondoz: Digital Speech, *Coding for low bit rate communication systems*, Wiley, 2004

[6] M. Jelinek, J. P. Adoul, *Frequency-Domain Spectral Envelope Estimation for Low Rate Coding of Speech*, ICASSP, 1999.

[7] P. Veprek, M. S. Scordilis, *Analysis, enhancement, and evaluation of five pitch determination techniques*, Speech Communication 37, 2002

[8] F. Charpentier, *Traitement de la parole par Analyse / Synthèse de Fourier application à la synthèse par diphones*, Thèse, ENST, Paris, 1988

[9] N. Henrich, B. Doval, C. d'Alessandro, and M.Castellengo. *Open quotient measurements on EGG, speech and singing signals*. In Proc. 4th International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research, Jena, Apr. 2000.

[10] G. Peeters. *Analyse et synthèse des sons musicaux par la méthode PSOLA*. In JIM98- Workshop, Agelonde, France, Mai 1998.

[11] A. Mousa. *Voice Conversion Using Pitch Shifting Algorithm by Time Stretching with PSOLA and Re-Sampling*. Journal of Electrical Engineering, Vol 61, NO. 1, 2010, p. 57-61

[12] P. Dutilleux, G. De Poli, U. Zölzer, *DAFX - Digital Audio Effects*, U. Zölzer, Ed. John Wiley & Sons, Sussex, England, 2002, p. 201-234.

[13] D. Childers, *Modern Spectrum Analysis*, IEEE Press, Piscateway, New Jersey, U.S., 1978, p. 252-255.

[14] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in C: the art of scientific computing*, Second Edition, Cambridge University Press, 1992.