J Multimodal User Interfaces (2008) 2: 133–144
DOI 10.1007/s12193-008-0010-4

ORIGINAL PAPER

RAMCESS 2.X framework—expressive voice analysis for realtime and accurate synthesis of singing

Nicolas d'Alessandro · Onur Babacan · Baris Bozkurt · Thomas Dubuisson · Andre Holzapfel · Loic Kessous · Alexis Moinet · Maxime Vlieghe

Received: 3 January 2008 / Accepted: 28 April 2008 / Published online: 5 June 2008

Abstract In this paper we present the work that has been achieved in the context of the second version of the RAMCESS singing synthesis framework. The main improvement of this study is the integration of new algorithms for expressive voice analysis, especially the separation of the glottal source and the vocal tract. Realtime synthesis modules have also been refined. These elements have been integrated in an existing digital instrument: the HANDSKETCH 1.X, a bimanual controller. Moreover this digital instrument is compared to existing systems.

N. d'Alessandro (✉) · T. Dubuisson · A. Moinet · M. Vlieghe
Circuit Theory & Signal Processing Laboratory, Faculté Polytechnique, Mons, Belgium
e-mail: nicolas.dalessandro@fpms.ac.be

T. Dubuisson
e-mail: thomas.dubuisson@fpms.ac.be

A. Moinet
e-mail: alexis.moinet@fpms.ac.be

M. Vlieghe
e-mail: maxime.vlieghe@fpms.ac.be

O. Babacan · B. Bozkurt
Electrical and Electronics Engineering Dpt, Izmir Institute of Technology, Izmir, Turkey

O. Babacan
e-mail: onur.babacan@iyte.edu.tr

B. Bozkurt
e-mail: baris.bozkurt@iyte.edu.tr

A. Holzapfel
Computer Science Dpt, University of Crete, Heraklion, Greece
e-mail: hannover@csd.uoc.gr

L. Kessous
LIMSI-CNRS, Université Paris XI, Paris, France
e-mail: loic.kessous@limsi.fr

Keywords Speech processing · Glottal source estimation · Realtime singing synthesis · Digital instrument design

1 Introduction

Expressivity is nowadays one of the most challenging topics studied by researchers in speech processing. Indeed recent synthesizers provide acceptable speech in term of intelligibility and naturalness but the need to improve human/computer interactions has brought researchers to developing systems that present more human-like expressive skills.

Expressive speech synthesis research seems to converge towards applications where multiple databases are recorded, corresponding to a certain number of labelled expressions. At synthesis time, the expression of the virtual speaker is set by choosing the units in the corresponding database, then the well-known unit selection algorithms are applied.

Recently remarkable achievements have also been reached in singing voice synthesis. The algorithms proposed by Bonada et al. [1] provide naturalness and flexibility by organizing singing frames at a high level. We can also highlight singing synthesis derived from STRAIGHT [2]. These kinds of technologies seem mature enough to allow the replacement of human singing with synthetic, at least for backing vocals.

However existing singing systems suffer from two restrictions: they aim at mimicking singers rather than offering creative voice timbre ability, and they are generally limited to note-based interaction, supposing the use of MIDI controllers.

In this context, we propose to investigate a novel approach. We postulate that, even if the use of databases is strategic in order to preserve naturalness, voice modeling

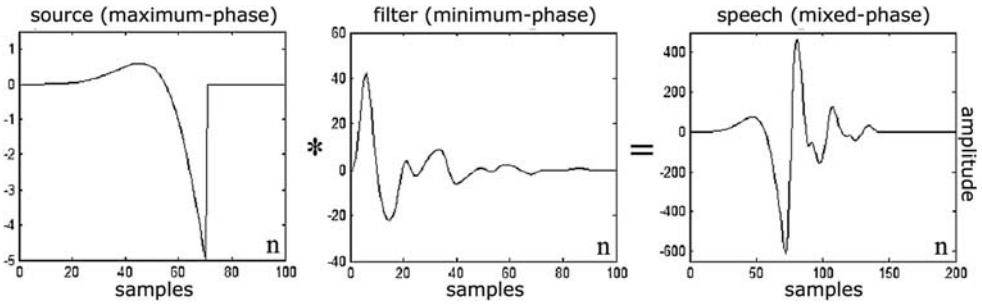

Fig. 1 Mixed-phase representation of speech: convolution of a maximum-phase source with a minimum-phase filter

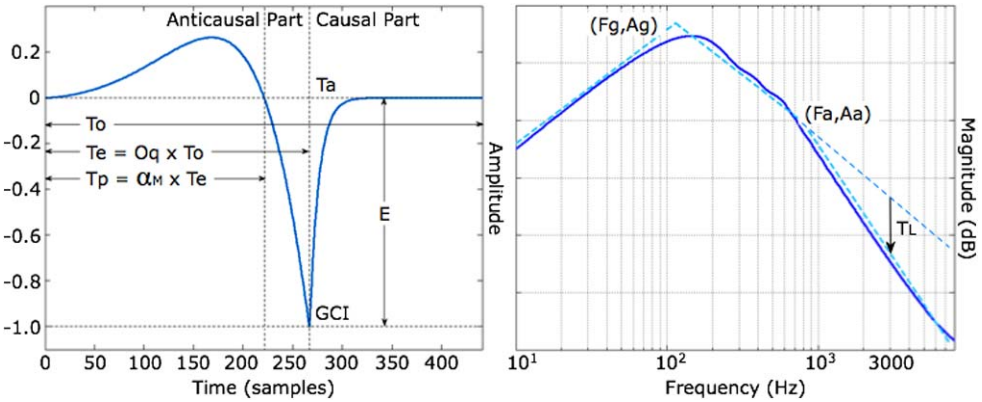

Fig. 2 Temporal and spectral representations of the GFD. Parameters are illustrated: T_0, $T_e = O_q T_0$, $T_p = \alpha_M T_e$, E, GCI in time domain, F_g, A_g, F_a, A_a, T_L in spectral domain

has to reach a more expressive level. These improvements have to meet particular needs, such as more realistic glottal source/vocal tract estimation, manipulation of voice quality features at a perceptual and performance level, and strong realtime abilities.

The second restriction is related to mapping strategies that have to be implemented in order to optimize the performer-synthesizer interface. Our aim is thus to create a coarticulated singing voice synthesizer which optimizes connections between voice quality aspects and the instrumental control.

In this paper we present the work that has been achieved in the context of the second version of the RAMCESS singing synthesis framework. This work has to be seen as an extension of existing reports in the field of eNTERFACE workshops [3], and related conference papers.

In Sect. 2 we introduce the mixed-phase model of speech production as the basis of our work. Then we give an overview of the database building (cf. Sect. 3). Sections 4 and 5 respectively describe analysis algorithms that have been used. Finally an overview of the RAMCESS 2.X software is presented, and a comparison with the state of the art is done.

2 Mixed-phase model of speech

In most of the speech processing literature Linear Prediction [4]—and thus minimum-phase—analysis is used as a basis of work. However recent investigations have proposed a mixed-phase speech model [5], based on the assumption that speech is produced by convolving an anticausal and stable source signal with a causal and stable vocal tract filter. The speech signal is thus a mixed-phase signal obtained by exciting a minimum-phase system (vocal tract) by a maximum-phase signal (glottal source). This convolution is illustrated in Fig. 1.

However, considering that the source is the anticausal part and the tract is the causal part is an approximation: a close observation of the behavior of vocal folds [6] shows us that the glottal flow (GF) waveform contains both a anticausal part (opened phase) and causal part (return phase). This aspect is even clearer on the glottal flow derivative (GFD) where the junction between anticausal and causal parts of the waveform creates a singular point called the glottal closure instant (GCI). A typical GFD period is illustrated in Fig. 2. Definition of temporal and spectral parameters of the glottal flow can be found in [7]. Using a mixed-phase model is equivalent with the assumption that the speech signal has two types of resonances. On the one hand causal resonances due to vocal tract acoustics, called formants. On the other hand an anticausal resonance called the glottal formant[1] [8].

In our study the mixed-phase modeling plays an important role in both analysis and synthesis. During analy-

[1] In the context of glottal source signals the term "resonance" or "formant" is used in order to emphasize the typical band-pass shape of the spectrum. From the acoustical point of view the GF is not a resonance.

sis, estimation of glottal flow is achieved by an algorithm (see Sect. 4.2) which decomposes the signal into anti-causal and causal components. During synthesis, mixed-phase signals are produced by exciting minimum-phase filters with maximum-phase—or mixed-phase—excitation (see Sect. 6.2). As a result we achieve a natural timbre for the output sound (respecting both magnitude and phase information of real acoustic signals). Moreover usual differences between speech and singing production can slightly be reduced. Indeed as we can access precisely to physiological parameters of the glottal source and the vocal tract, we can simulate some of these differences: glottal source registers, singing formant [9], larynx lowering, etc.

3 Database preparation

In order to achieve the synthesis of coarticulated speech we need to record, segment and pre-process a set of waveforms (that we call our *database*) containing instances of expected phonetic sequences. In this section we discuss choices made in the design and the segmentation of the database. As we plan to achieve a GCI-synchronous analysis (analysis window centered on GCI of each period) an algorithm used to achieve a GCI labeling is also presented.

3.1 Corpus design

The development of a fully functional synthesizer, able to pronounce any sentence requires the recording of—at least—one instance of each diphone (i.e. a two-phoneme sequence) of a language, in order to preserve coarticulation [10]. This step is particularly time consuming. As our purpose is the proof of concept of an analysis/resynthesis method based on mixed-phase speech modeling, we only target the synthesis of a small amount of sentences. Therefore, it is not necessary to record a full corpus and restricting to a small set of diphones is sufficient.

Database is prepared as follows. The subject is asked to pronounce words, made of arbitrary sequences of phonemes. We used sequences of voiced-unvoiced phonemes in order to avoid the problem of voiced transients (e.g. /*b*/) analysis. A typical word that has been used is /*t a k a p a f a S e*/ (in SAMPA notation [11]). Each sentence is recorded with constant pitch and constant voice quality (as much as possible), so as to constrain the range of subsequent analysis steps.

With this method, the voice recorded in the database consists of words pronounced in a "singing style" (long vowels, none lyric), with no vibrato and a constant pitch. Besides, as the speaker is not a trained singer, the behavior of the glottal source is close to speech. As a result the database content is neither natural speech nor natural singing but somewhere in-between. The main motivation behind this is simplification of the analysis process. Indeed the models and methods used in analysis (see Sects. 4 and 5) have been developed mainly for speech analysis. Moreover considering various styles of singing and building an analysis method for such a database would push us to a very large scale research. Therefore we decided to limit our work in this sense and concentrate on other details using our limited resources.

As the constancy of the pitch is difficult to obtain in a free recording situation, a synthesized reference (constant pitch/quality) is played first, and then the speaker's imitation is recorded. Each sentence is recorded twice, at different speeds in order to vary the coarticulation conditions. Voice is recorded at a sample rate (F_s) of 44.1 kHz and with 16 bits of resolution.

3.2 Database segmentation

The phonetic sequences are manually segmented into two sets: the vowels and the consonants. Each vowel is in turn divided into three parts. The first and third parts are the regions of the vowel that are coarticulated respectively with the preceding and following consonant (or silence). These are the left and right transient parts of the vowel. These coarticulated parts of speech usually contain a few (less than ten) periods of the vowel, depending on the average fundamental frequency. The central part, also referred as the steady state part is the region of speech considered as coarticulation-free and thus actually quasi-stationary. An example is illustrated in the Fig. 3.

3.3 Initial GCI marking

With the samples segmented as described in Sect. 3.2, the voiced parts can be processed separately in order to get a first estimation of the GCI positions. Considering our corpus, this task is relatively straightforward. Indeed speech signals only contain sequences of unvoiced consonants and vowels (e.g. /*S e*/). Thus the onset of the voiced island is clearly visible. This onset is the first glottal pulse of the phonation, not yet added to previous vocal tract impulse response. Practically this means that the first GCI of the island—that we propose to call GCI_1—can be located by a negative peak search around the segmentation point. The searching technique is described in (1).

$$GCI_1 = \min_{n=[L,L+T_0]} x(n) \tag{1}$$

where $x(n)$ is the signal of the database, T_0 is the local fundamental period (estimated e.g. by autocorrelation), and L is the segmentation point starting the current voiced island.

From the position of GCI_1 and an estimation of the fundamental frequency along the signal, other GCI positions

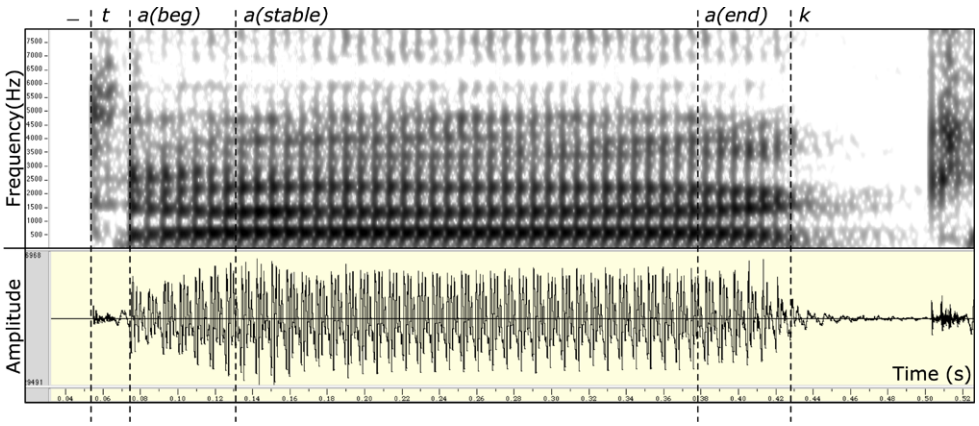

Fig. 3 Typical example of segmentation. Consonants and vowels are separated, and inside a vowel, begin and end of the coarticulation process are identified

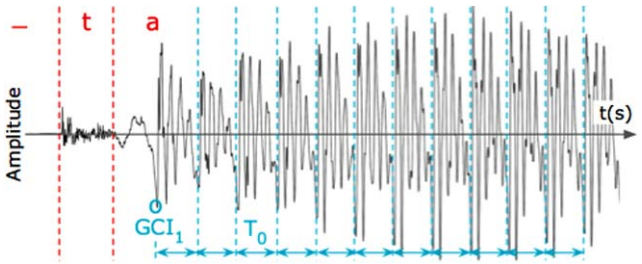

Fig. 4 GCI_1 is located slightly after the segmentation point, and GCI_k positions are extrapolated from locally estimated T_0

can be extrapolated. These marks are referenced as GCI_k, in the considered island. We work in a GCI-synchronous framework. This means that the kth analysed frame is centered on GCI_k. Moreover the window length is set at $2 \times T_0$ (two times the local fundamental period).

We also introduce a particular notation. $x_{V,k}$ is the *Voice* frame (signal from the database, without any particular processing) extracted around the kth GCI.

4 Anticausal part estimation

Considering the mixed-phase model of speech, we can emphasize that causality is a discriminant factor in order to isolate a part of the glottal source signal (the open phase). In this section we describe the algorithm used in order to achieve a first separation of anticausal and causal contributions, using Zeros of the Z-Transform (ZZT) [12]. Then we show results in order to highlight some typical problems of straightforward decomposition (as exposed in [13]). We propose a method to improve these decomposition results, based on the adjustment of GCI positions. Finally we describe a time-domain fitting technique in order to extract some relevant GFD parameters (O_q, α_M).

4.1 Zeros of the Z-transform (ZZT)

For a series of N samples $(x(0), x(1), \ldots, x(N-1))$ taken from a discrete signal $x(n)$, the ZZT representation (Zeros of the Z-Transform) is defined as the set of roots (zeros of the polynomial) $\{Z_1, Z_2, \ldots, Z_m\}$ of its corresponding Z-Transform $X(z)$, as illustrated in (2).

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} = x(0) z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m) \tag{2}$$

This representation implies to compute roots of polynomials [14] of which the degree increases with respect to the sampling frequency. It introduces errors on the estimation of zeros in high frequencies, due to the iterative computation of roots. For this reason the ZZT computation is usually performed at 16 kHz. It means that our database has to be downsampled.

4.2 ZZT-based causal/anticausal separation

According to the mixed-phase model of speech, the ZZT representation of a speech frame contains zeros due to the anticausal component and to the causal component [15]. Consequently zeros due to the anticausal component lie outside the unit circle, and zeros due to the causal component inside the unit circle. Under some conditions about the location, the size and the shape of the analysis window, zeros corresponding to both anticausal and causal contributions can be properly separated by sorting them out according to their radius in the z-plane, as illustrated in Fig. 5.

The waveform and the spectrum of these contributions are then computed by the Discrete Fourier Transform (DFT) presented in (3).

$$X(e^{j\phi}) = G e^{(j\varphi)(-N+1)} \prod_{m=1}^{N-1} (e^{(j\varphi)} - Z_m) \tag{3}$$

Applying this algorithm on our database, we identify *Causal* and *Anticausal* frames resulting from the ZZT-based decomposition around the kth GCI, respectively by $x_{C,k}$ and $x_{A,k}$. Examples if these frames are illustrated in Fig. 6.

4.3 Re-adjustment of GCI positions

Bozkurt has shown that the location, the size and the shape of the analysis window have a significant impact on the effectiveness of the ZZT-based decomposition, i.e the noisiness of $x_{C,k}$ and $x_{A,k}$ components [5]. We can highlight that the centering of the window around the GCI has to be really precise, as the decomposition algorithm is really sensitive to wrong positioning. In the context of this study, we first use this sensivity in order to refine location of GCI_k, as instants where the decomposition is properly achieved.

The improvement of GCI position is obtained by the combination of two mechanisms [13]. On the one hand, systematic shifts are realized around each GCI_k. If the maximum shift range is set e.g. to 4 samples, 9 $x_{A,k}$ candidates are computed around each GCI_k. On the other hand, a criterion is computed by comparing $x_{A,k}$ with a "proper anticausal component". This measurement is made in spectral domain. Indeed by comparing magnitude spectrum of a correct $x_{A,k}$ (close to the anticausal component of the glottal flow) and a noisy $x_{A,k}$, we can observe that their behaviour is quite similar below 2 kHz, but significantly different in higher frequencies (see Fig. 7).

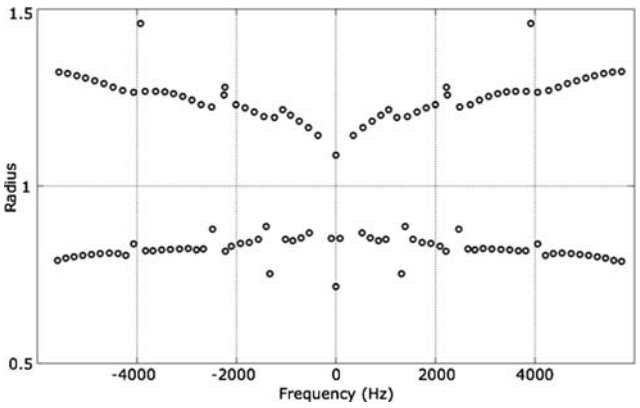

Fig. 5 Distribution of Z_m in the Z plane (polar coordinates), showing that inner and outer zeros can be sorted out. This example results from on a synthetic mixed-phase frame

In order to choose the best $x_{A,k}$ among all candidates (for a given k), we define our spectral criterion as the ratio between the energy in the 0–2 kHz frequency band and the energy in the whole spectrum (0-$F_s/2$), as presented in (4).

$$C = \frac{Energy_{[0-2000\,\mathrm{Hz}]}}{Energy_{[0-8000\,\mathrm{Hz}]}} \tag{4}$$

For each GCI_k, the best $x_{A,k}$ is chosen as the one which maximizes this criterion among all the candidates. The temporal location of each GCI_k can thus be refined, as well as the estimation of the fundamental frequency (f_0).

4.4 O_q and α_M estimation

Once the best possible $x_{A,k}$ have been obtained for each k, interesting glottal flow parameters can be estimated on these frames: open quotient (O_q) and asymmetry coefficient (α_M). For a given k, these parameters result from a time domain fitting between $x_{A,k}$ and a typical Liljencrants-Fant (LF) glottal pulse model (opened phase only) [16]. Our implementation of this model is described in Sect. 6.2. The fitting strategy is realized given the following steps:

- current period (T_0) and negative peak amplitude (E) is extracted from $x_{A,k}$;
- ranges of variation and resolutions of O_q and α_M are decided. For example, O_q can vary in $[0.3 - 0.9]$ and α_M in $[0.6 - 0.9]$, both by steps of 0.05; These values are represented by $\vec{O}_q$ and $\vec{\alpha}_M$ vectors;
- a codebook Θ_F, containing an array of LF pulses (opened phase only), is computed based on T_0, E, $\vec{O}_q$ and $\vec{\alpha}_M$;
- a mean square error between each instance of Θ_F and $x_{A,k}$ is computed, resulting in an array of error values Err_k;
- the smallest element of Err_k indicates which element of Θ_F fits $x_{A,k}$ the best (in the sense of MSE), and thus providing values for O_q and α_M.

This procedure is formally presented in (5)–(8). The fitting algorithm also results in a LF frame, close to $x_{A,k}$. This

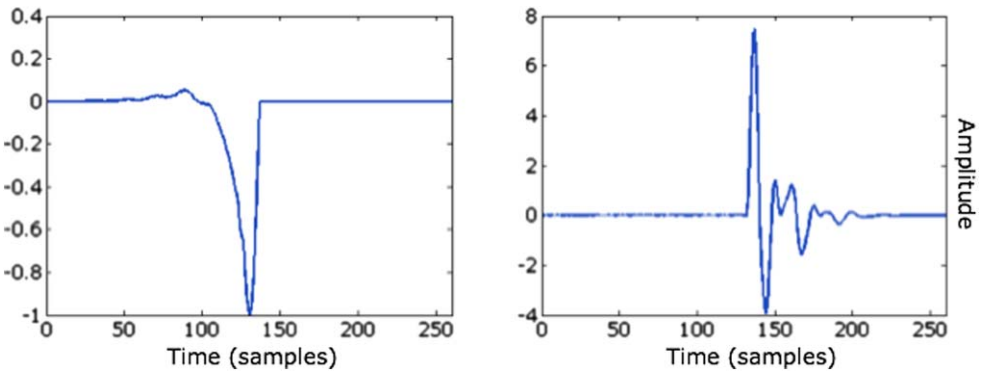

Fig. 6 Effective results of the ZZT-based decomposition on a real speech frame of /a/: $x_{C,k}$ (*right*) and $x_{A,k}$ (*left*)

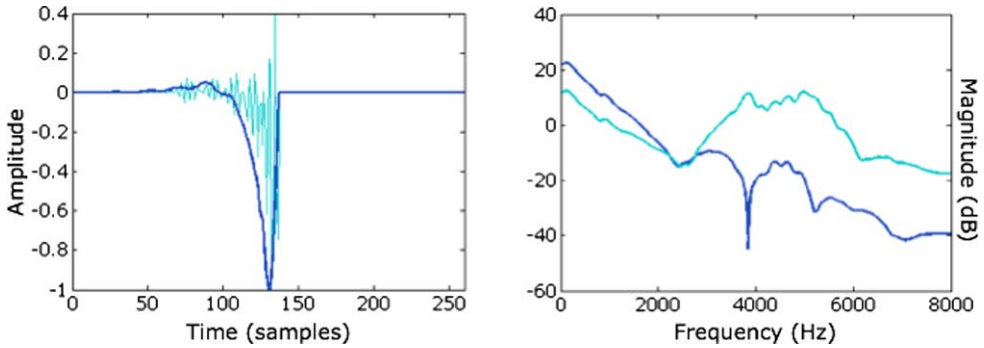

Fig. 7 Good $x_{A,k}$ (*dark*) vs wrong $x_{A,k}$ (*light*): illustration of the spectral criterion based on the clear increasing of high frequencies when decomposition fails

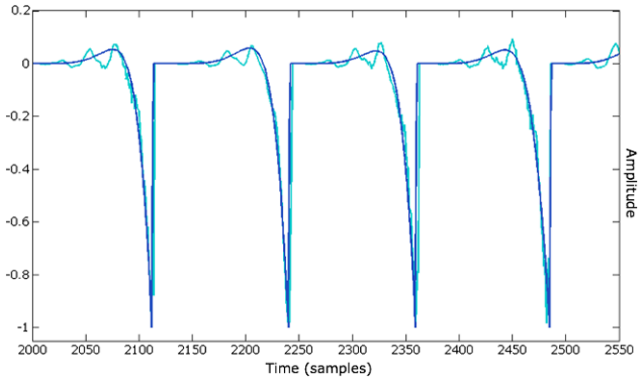

Fig. 8 Fitting in time domain between the anticausal component coming from ZZT-based decomposition $x_{A,k}$ (*light*) and the LF model of glottal pulse $x_{F,k}$ (*dark*)

Fitted frame is referenced as $x_{F,k}(n)$ (see (9)).

$$\Theta_F = f(T_0, E, \vec{O_q}, \vec{\alpha_M}) \tag{5}$$

$$Err_k(m,n) = (\Theta_F(m,n) - x_{A,k})^2 \tag{6}$$

$$Err_k(a,b) = \min_{m,n} Err_k(m,n) \tag{7}$$

$$O_q = \vec{O_q}(a), \qquad \alpha_M = \vec{\alpha_M}(b) \tag{8}$$

$$x_{F,k} = \Theta_F(a,b) \tag{9}$$

In Fig. 8 we present results of the fitting algorithm on several $x_{A,k}$ periods. It shows that estimated O_q and α_M values are coherent ans stable, which is expected as the database has been recorded with constant voice quality. This assumption can also be verified by observing O_q and α_M values (see Fig. 9).

Moreover an evaluation of the fitting error on a larger database shows that the process is stable and the error remains low (around 2.5% of the intensity of the signal) on typical kinds of voiced phonemes. The fitting error for the kth frame (which has a length $= N$) is measured with by (10) and the result on 500 frames is illustrated in the Fig. 10.

$$Err_{fit}(k) = \frac{1}{N} \sum_{n=0}^{N-1} |x_{F,k}(n) - x_{A,k}(n)| \tag{10}$$

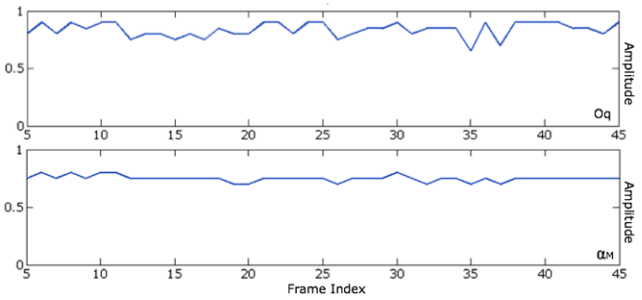

Fig. 9 Evolution of estimated open quotient (O_q) (*upper panel*) and asymmetry coefficient (α_M) (*lower panel*), coming from the time-domain fitting, for a /a/ sound in the database

5 Causal parts estimation

The use of the ZZT-based algorithm provides a first estimation of the evolution of O_q and α_M (parameters of the anticausal part) along the database. Once these values are estimated they are used as inputs in order to determine other components of speech frames: return phase shape (through spectral tilt T_L) and auto-regressive coefficients of the vocal tract. In this section we describe the use of the ARX-LF algorithm in the jointly estimation of the complete glottal pulse (opened and return phase) $x_{G,k}$ and the vocal tract impulse response ($x_{T,k}$ then $x_{U,k}$). The combination of both ZZT and ARX-LF algorithms, focused on their respective strenghts, finally produces a better estimation.

5.1 Auto-regressive eXogeneous model (ARX)

In the source/filter model [17], a sample $s(n)$ of speech is modeled by the auto-regressive (AR) equation:

$$s(n) = -\sum_{i=1}^{p} (a_n(i)s(n-i)) + b_n u(n) + e(n) \tag{11}$$

where $u(n)$ and $e(n)$ are samples of the source and the residual and $a_n(i)$ and b_n are the AR filter coefficients (these coefficients represent the time-varying vocal tract).

Contrary to LP analysis, Vincent et al. [18] and [19] assume that the source u is a glottal flow in the voiced part

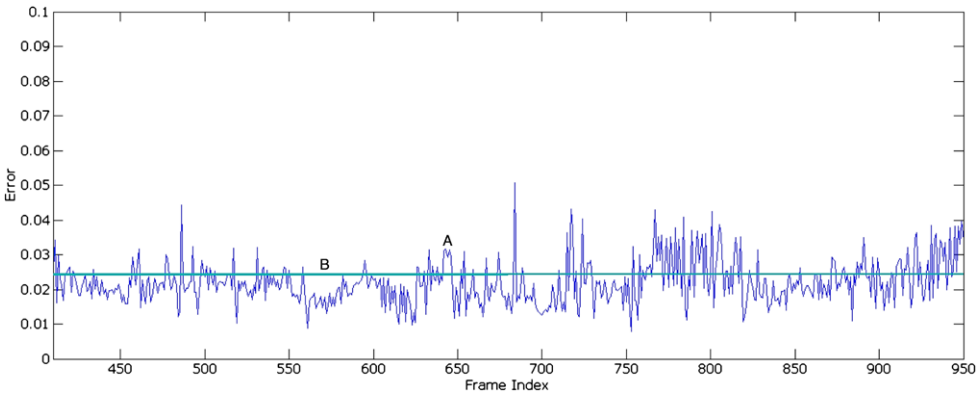

Fig. 10 Computation of the fitting error, estimated by (10), between $x_{F,k}$ and $x_{A,k}$, on 500 frames (A) and the mean value (B)

of speech. However this change prevents the use of Yule-Walker equations. Instead one has to solve the system of linear equations (12) obtained by writing (11) for successive values of n.

$$S = MA + E \tag{12}$$

where S is a vector of (possibly windowed) speech samples $s(n)$, M is the concatenation of a matrix of $-s(n-i)$ values and a vector of glottal source samples $u(n)$. A is the vector of unknown values $a_k(i)$ and b_k. E is a vector of residual samples $e(n)$, that is to say a vector of modeling errors that we want to minimize when computing A. Values of n are chosen with respect to the framework defined in Sect. 3.3 (i.e. GCI-synchronous and $2 \times T_0$ window length: $n = [GCI_k - T_0 + 1 \dots GCI_k + T_0]$).

Finding the unknown values $a_k(i)$ and b_k requires to define a set of glottal sources $\Theta = [u_1 \dots u_W]$ and to choose among these the one which minimizes the modeling error of the ARX model. That is to say it requires to solve the system of equations for each u_w and selecting the one that minimizes $\|E\|^2$. That glottal flow u_w minimizing the modeling error is considered as the most accurate estimate of the actual glottal flow produced by the speaker. Likewise the resulting parameters $a_k(i)$ and b_k are estimate of the actual speaker's vocal tract.

5.2 ARX-LF optimization on a sub-codebook

In order to determine the causal parts (i.e. the vocal tract parameters and the return phase, through the spectral tilt value T_L) of each frame $x_{V,k}$ we use a modified version of the ARX method described in the previous section.

Indeed a complete codebook Θ of glottal flows based on the possible variations of their parameters (O_q, α_M and T_L in our implementation, as explained in Sect. 6.2) would be rather bulky and solving (12) for all the words of that codebook would be computationally expensive. Moreover it has been highlighted that ARX-LF could sometimes converge to very unprobable consecutive values, making it unstable if used alone.

Fortunately O_q and α_M estimations are already known for each GCI_k (thanks to ZZT analysis and LF fitting techniques) which allows us to select a part of the codebook Θ that we call the sub-codebook $\Theta_{S,k}$. T_L is the only varying parameter of that sub-space. Moreover although we are confident in the estimate of O_q and α_M described in Sect. 4.4, we could refine these results by selecting a somehow larger sub-codebook, allowing slight variations of O_q and α_M around their initial estimations.

For each GCI_k, the corresponding sub-codebook $\Theta_{S,k}$ contains W glottal flows. We compute the $a_k(i)$ and b_k coefficients for every word of $\Theta_{S,k}$ and then re-synthesize an approximation $x_{w,k}$ of the frame of speech $x_{V,k}$ by applying (11). At GCI_k, the error for the wth synthetic frame $x_{w,k}$ is then measured as its Euclidean distance to the original frame $x_{V,k}$.[2]

$$E_w = \sqrt{\sum_{n=1}^{2T_0} (x_{V,k}(n) - x_{w,k}(n))^2} \tag{13}$$

However, before actually computing the errors, two important points remain: the position of GCI_k and the stabilization of the AR filter. Indeed, the estimate of each GCI position is provided by the ZZT analysis. Although that position fits very well for ZZT decomposition, it's not necessarily the best one for ARX decomposition. For that reason one more step is added to the algorithm explained above: just like during ZZT analysis we do not consider only the analysis window $x_{V,k}$ centered on GCI_k but also windows centered a few points on the left and on the right of that location.

In our implementation we look three samples before and after the position of GCI_k. Henceforth we will have $7 \times W$ $x_{w,k}$ and corresponding error measurements. Not only the minimum error will give us $x_{G,k}$, the best guess for the glottal flow (with O_q, α_M and T_L as its parameters), but it also provide us with the optimal position of GCI_k.

[2] Both $x_{w,k}$ and $x_{V,k}$ are hanning-windowed.

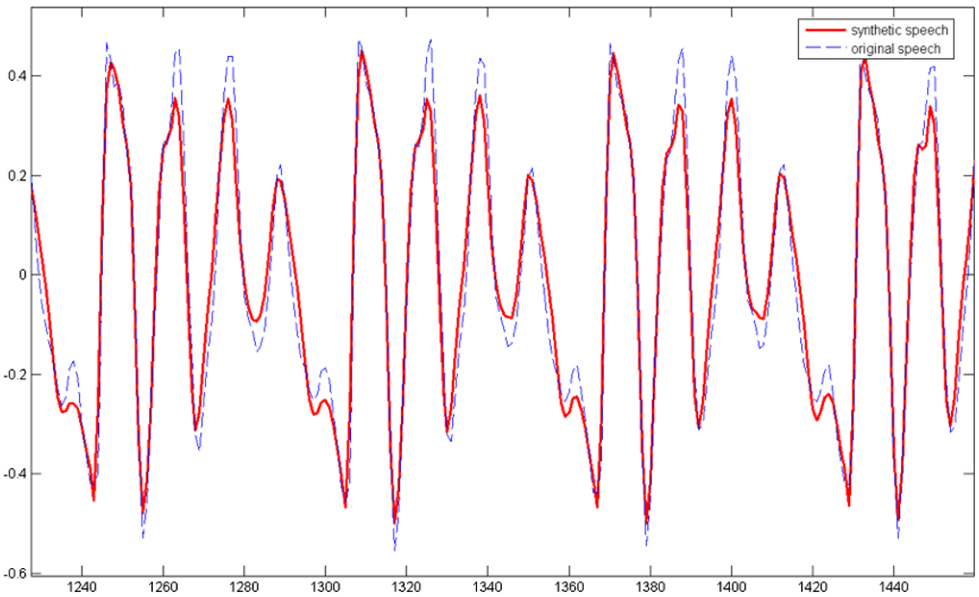

Fig. 11 Superposition of original (*dashed*) and resynthesized (*solid*) signals, after the computation of ARX-LF on a sub-codebook dened by ZZT-based parameters

Finally, although LP analysis guarantees that the AR filter has all of its poles inside the unit circle and therefore is stable, this is no longer the case when solving (12). Consequently, the last step before synthesizing any of the $x_{w,k}$ is to reflect the outside poles of a_k inside the unit circle and adapting the value of parameter b_k.

All these steps are performed at a sample rate of 8 kHz which allows us to get reliable estimations of T_L and the positions of GCI_k (as well as an estimate of the filter parameters, a_k and b_k, for that rate). However high quality singing synthesis is produced at higher sampling rate such as 44.1 kHz.

The glottal flow parameters O_q, α_M and T_L are independent of the sampling frequency and therefore they can be used as is. On the contrary the filter coefficients rely upon the sampling rate and need to be recomputed. The task is fast and easy since we just have to solve (12) once with a different sampling frequency with $x_{V,k}$ (for which we have the original recording at 44.1 kHz) as S and $x_{G,k}$ (which is a parametric model and thus can be produced at any given rate) as u.

Equation (12) is first solved at 8 kHz for 24 $a(i)$ parameters ($p = 24$) and considering a sub-codebook with O_q and α_M constant and T_L varying between 3 dB and 20 dB (with a 1 dB step). Then it is solved at 44.1 kHz for $p = 46$ and O_q, α_M and T_L constant. The vocal tract parameters a_k and b_k thus obtained define a filter with $x_{T,k}$ as its impulse response. An example of resynthesis for a few frames is presented in Fig. 11.

5.3 Vocal tract filter compensation

It's observed that synthesis obtained by exciting the ARX filter with the glottal flow results in a certain loss of high frequency components. This effect is due to the ARX optimization, focusing on a time-domain reduction of the error, thus first encouraging low frequencies fitting. To compensate for this effect, we devised a simple compensation method via AR filtering. For this, the AR compensation filter is obtained by linear predictive analysis of an impulse response obtained in the following way. The frequency response of the original signal is divided by the frequency response of the synthetic signal, and the inverse Fourier transform of the result is taken. A sample result of the compensation is presented in Fig. 12. The obtained AR compensation filter is combined (by cascading) with the ARX filter to obtain a single filter that will perform the synthesis in one stage. This *Updated* vocal tract filter is referenced as $x_{U,k}$.

5.4 Overall evaluation of the analysis process

As a first quantitative evaluation of the whole analysis pipeline, we compute a distance for each frame k (which has a length $= N$) between $x_{V,k}$ (the original signal) and $x_{U,k}$ (the final resynthesis). The overall error is computed by (14) and the result is illustrated for 500 frames in the Fig. 13. It can be highlighted that for most of the frames the error remains low (around 5% of the intensity of the signal). Moreover peaks appear for some frames which make the link between voiced and unvoiced regions and present mixed harmonic/noise properties.

$$Err_{all}(k) = \frac{1}{N}\sum_{n=0}^{N-1} |x_{U,k}(n) - x_{V,k}(n)| \tag{14}$$

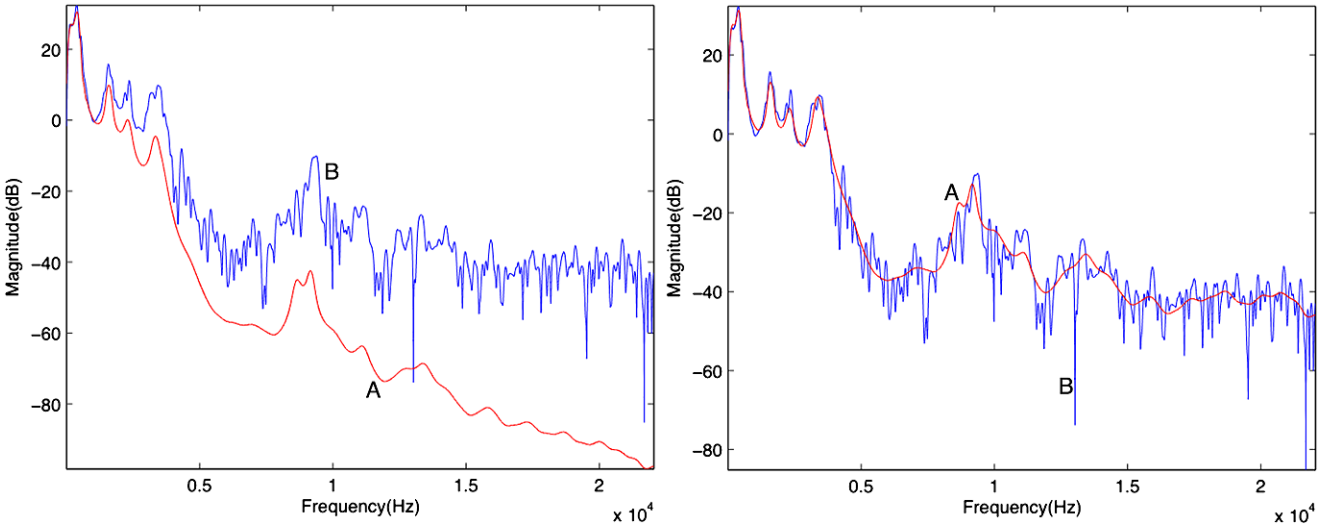

Fig. 12 Spectral envelope of original (B) and synthetic (A) signals before (*left*) and after (*right*) the HF compensation

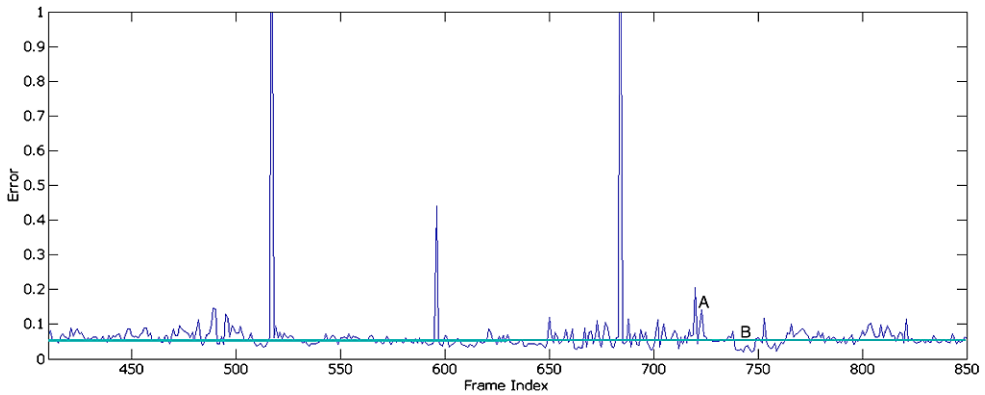

Fig. 13 Computation of the overall error, estimated by (14), between $x_{U,k}(n)$ and $x_{V,k}$, on 500 frames (A) and the mean value (B)

6 RAMCESS 2.X framework

In this section we first explain how the different algorithms—described in previous parts—have been gathered together in order to implement a functional MATLAB analysis pipeline. Then, as our underlying purpose is the achievement of the second version of the RAMCESS software, we present the main improvements from the RAMCESS 1.X. The RAMCESS project aims at proposing a fully functionnal singing voice synthesizer, through a shared library, working in realtime in music programming environments such as Max/MSP [20] or Pd [21]. Then the integration of the last release of RAMCESS inside an existing digital instrument—the HandSketch [22]—is described. Finally a rating of the system, compared to the state of the art, is presented.

6.1 The analysis pipeline

Our analysis pipeline is a library of MATLAB functions (*XsLib*). This consecutively applies the framing of the recorded sound files ($x_{V,k}$), the ZZT-based decomposition ($x_{A,k}$, $x_{C,k}$), the time domain fitting ($x_{F,k}$), the ARX-LF optimization ($x_{G,k}$, $x_{T,k}$), the vocal tract compensation ($x_{U,k}$), and finally the encoding of parameters of the mixed-phase model in a SDIF file [23]. These different steps are illustrated in Fig. 14.

6.2 Spectrally-enhanced LF model (SELF)

As we assume that speech signal has mixed-phase properties, the RAMCESS 2.X synthesis engine is built on a flexible and realtime GF/GFD generator. Glottal pulse can be synthesized with many different ways. In term of flexibility and quality, we can particularly highlight LF [16] and CALM [8] models. However none of them can be transposed directly in the realtime context. If we observe LF equations of the GFD $g'(t)$:

$$g'(t) = \begin{cases} -Ee^{a(t-T_e)}\frac{\sin(\pi t/T_p)}{\sin(\pi T_e/T_p)}, & 0 \le t \le T_e, \\ \frac{-E}{\epsilon T_a}(e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_0-T_e)}), & T_e \le t \le T_0 \end{cases} \quad (15)$$

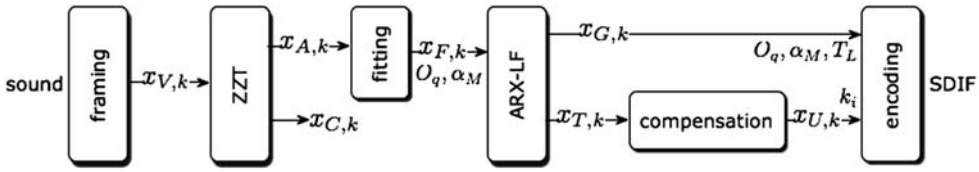

Fig. 14 Overall illustration of the steps of the analysis pipeline, showing the different frames involved in the process

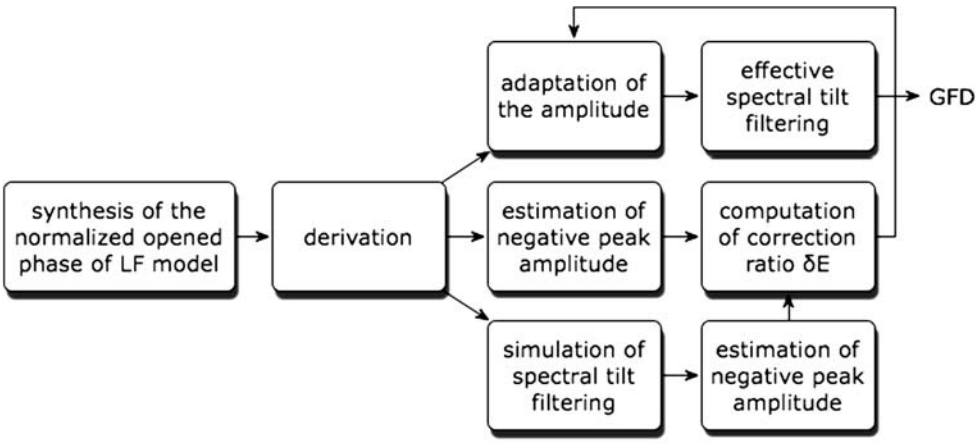

Fig. 15 Different steps in the synthesis of the GFD in the SELF engine: a first simulation of the spectral tilt filtering is performed in order to compute the amplitude correction

where T_a, T_p, T_0, E can be identified in the Fig. 2, we notice a and ϵ are scaling parameters. A system of two implicit equations[3] (see (16)) has to be solved in order to determine these parameters of LF curves in (15). Typical implementations of LF pulse synthesis have shown that such a system is known to be unstable in a given range of (O_q, α_M) variation.

$$\begin{cases} \epsilon T_a = 1 - e^{-\epsilon(T_0 - T_e)}, \\ \frac{1}{a^2 + (\frac{\pi}{T_p})^2} \left(e^{-aT_e} \frac{\pi/T_p}{\sin(\pi T_e/T_p)} + a \right. \\ \quad \left. - \frac{\pi}{T_p} \operatorname{cotg}(\pi T_e/T_p) \right) = \frac{T_0 - T_e}{e^{\epsilon(T_0 - T_e)} - 1} - \frac{1}{\epsilon} \end{cases} \tag{16}$$

Concerning CALM this problem is solved by using linear filters for the computation of GF/GFD waveforms, but one of these filters have to be used anticausally. This operation is possible in realtime but with a limited flexibility [24].

The improvement that we propose can be seen as a compromise between both LF and CALM models. We propose to call it the *Spectrally-Enhanced LF model* (SELF). In order to avoid the solving of the system of implicit equations, only the left part of the LF model is used. It is computed using the left part of the normalized glottal flow model [7] (see (17)).

$$n_g(t) = \frac{1 + e^{at}(a\frac{\alpha_m}{\pi} \sin(\pi t/\alpha_m) - \cos(\pi t/\alpha_m))}{1 + e^{a\alpha_m}} \tag{17}$$

[3]The LF model gives equations of temporal shapes of both curves on the left and the right of the GCI. The conditions are then (1) the integration of the whole period has to be 0, and (2) left and right curves have to be connected at the position of the GCI.

where t evolves between 0 and 1, and is sampled in order to generate the $O_q \times \frac{F_s}{F_0}$ duration of the opened phase ($F_0 = 1/T_0$, the fundamental frequency) and $a = f(\alpha_m)$ is the preprocessed buffer of solutions of (18).

$$1 + e^a \left(a \frac{\alpha_m}{\pi} \sin\left(\frac{\pi}{\alpha_m} - \cos\left(\frac{\pi}{\alpha_m} \right) \right) \right) \tag{18}$$

Then the right part (the return phase) is generated in spectral domain, which means that the left LF pulse is filtered by the spectral tilt low-pass first order filter (controlled by T_L) presented in [8]. This option is also preferred because a long filter-generated return phase smoothly overlaps with following pulses, thus avoiding discontinuities.

The full algorithm also integrates the derivation of the pulse and a normalization of E, in order to control separately spectral tilt and energy of the signal. Moreover it has been noticed by experiment that the realtime normalization can not be realized after the spectral tilt operation. It creates discontinuities. Thus these two steps have to be interchanged. Consequently a first measurement of the impact of spectral tilt filtering on E has to be done. δE is defined as the ratio between E after and before spectral tilt filtering. Then the normalization process happens on $n_g(t)$ with a scaling factor of $1/\delta E$. Finally the signal is passed through the spectral tilt filter, and the resulting E is 1. The algorithm is presented in Fig. 15.

6.3 Dimensional control of voice quality

On top of our synthesis parameters (F_0, O_q, α_m, T_L, spectral and geometrical features of vocal tract), it is interesting

Fig. 16 A typical use of the HandSketch digital instrument

to build a layer which is able to provide more perception-based control to the performer. This dimensional study of voice quality has been achieved, resulting in a set of dimensions and their corresponding mapping equations with synthesis parameters [24]. In RAMCESS 2.X, *tenseness* (T), *vocal effort* (V) and *registers* (M_i) got a particular attention. Relations with O_q, α_M and T_L are presented in (19) to (24).

$$\begin{cases} O_q = O_{q_0} + \delta O_q, \\ \alpha_M = \alpha_{M_0} + \delta\alpha_M, \\ T_L = T_{L_0} \end{cases} \tag{19}$$

$$O_{q_0} = \begin{cases} 0.8 - 0.4\ V, & M_i = M_1 \text{ (modal)}, \\ 1 - 0.5\ V, & M_i = M_2 \text{ (chest)} \end{cases} \tag{20}$$

$$\alpha_{M_0} = \begin{cases} 0.8, & M_i = M_1 \text{ (modal)}, \\ 0.6, & M_i = M_2 \text{ (chest)} \end{cases} \tag{21}$$

$$T_{L_0}(dB) = 55 - 49V \tag{22}$$

$$\delta O_q = \begin{cases} (1 - 2\ T)\ O_{q_0} + 0.8\ T - 0.4, & T \leq 0.5, \\ (2\ T - 1)\ O_{q_0} + 2\ T + 1, & T > 0.5 \end{cases} \tag{23}$$

$$\delta\alpha_M = \begin{cases} (0.25 - 0.5T)\alpha_{M_0} + 0.4T - 0.2, & T \leq 0.5, \\ (0.5T - 1)\alpha_{M_0} - 1.2T + 0.6, & T > 0.5 \end{cases} \tag{24}$$

6.4 Integration in the HANDSKETCH 1.X instrument

In order to follow our hypothesis, expressive and interactive properties of a singing synthesis system have to be considered together. Thus the final attempt of this research is the integration of coarticulated speech ability into an existing digital instrument, the HANDSKETCH 1.X [22]. An illustration of the controller is presented in the Fig. 16. This instrument proposes a bi-manual configuration in order to control expressively the voice quality dimensions that have been described in Sect. 6.3. At this level our encoded database is the element to be loaded and interfaced in the performance application.

Browsing the SDIF file provides time-tagged values for F_0, O_q, α_M, T_L and k_i (reflection coefficients of the vocal tract). The existing HANDSKETCH application proposes mapping strategies for the dimensional control of source information: F_0, O_q, α_M and T_L. Consequently the integration of the database in the existing workflow corresponds to two different problems. On the one hand, the vocal tract can be controlled by one-to-one mapping with the database values. At this level the critical aspect is the triggering of phonetic information, which is really difficult to achieve manually. On the other hand, two different glottal source definitions have to be merged together: one from the dimensional control, and one from the database. A first proposition is simply to ignore database information, which gives acceptable results. The other option is that the dimensional control affects values of the database by deviation, which results in more natural synthesis, but is more difficult to control.

Table 1 Results of the votes of Interspeech 2007 SSC

	Art	*Str*	*Voc*	*For*	*Ram*
Expr	3.1	2.3	3.1	3.3	2.6
Over	2.8	2.4	3.1	3.5	2.9

6.5 Evaluation of the expressivity

In order to test the expressivity of our system, some choices have been made on the mapping: the glottal source is controled by the dimensional mapping (thus ignoring the database information), and the articulation between the phonemes is achieved by triggering movements in the database. As a way to evaluate the current system (RAMCESS 2.X + HANDSKETCH 1.X), we subscribed to a special session of Interspeech 2007 international conference, called "Synthesis of Singing Challenge" [25]. It has been an excellent opportunity to compare with significant contributors in this topic: Birkholtz's articulatory synthesis [26], STRAIGHT [2], Vocaloid [1], or Sundberg's system [27].

The Table 1 reports the votes concerning the expressivity (*Expr*) and the overall judgement (*Over*) for five systems: Birk-holtz's articulatory synthesis (*Art*), STRAIGHT (*Str*), Voca-loid (*Voc*), Sundberg's formant synthesis (*For*), and RAMCESS 2.X plus HANDSKETCH 1.X combination (*Ram*). Judgements were given by about 60 voters from the audience (researchers from speech processing community), with a five-point Likert scale from excellent (1) to poor (5). We can highlight that, concerning these two aspects, our system is second, just behind STRAIGHT, especially for expressivity. This point is particularly encouraging, as this

work was our first opportunity to nest the different analysis algorithms presented in this paper, and unify their results in a realtime and interactive framework.

7 Conclusion

In this paper we described a novel approach to singing synthesis that combines database concatenation synthesis techniques with source/filter synthesis. This method transforms a vowel-only system into a framework which is able to produce coarticulated speech expressively with high quality consonants as compared to a simple source/filter synthesis. This work illustrated that the deep analysis of glottal source features and the separation with vocal tract components were achievable on a prepared database. For this purpose the ZZT-decomposition method is effectively coupled with ARX modeling and successfully tested for the first time within an analysis-synthesis framework. Furthermore we were also able to validate the use of new realtime synthesis modules like SELF, and more generally RAMCESS elements in a complex voice production context. Finally the participation to a voting session on singing synthesis quality provided encouraging results, placing our system at the second place, compared to the state of the art. On the top of this current work, two significant axis will be developed further: the use of a larger database for unit selection, and the extension of these concepts to musical signal synthesis, as recent work have shown that it was clearly possible [28].

Acknowledgements Authors would first like to thank the organization committee of eNTERFACE07 summer workshop in Istanbul (Bogazici University) for their great involving and support. We also would like to highlight a role of the European Union (through FP6 and SIMILAR) in the achievement of all these tools. Finally we would like to thank our respective laboratories and supervisors for their advices, trust and support.

References

1. Bonada J, Serra X (2007) Synthesis of the singing voice by performance sampling and spectral models. IEEE Signal Process 24(2):67–79
2. Kawahara H (1999) Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds. Speech Commun 27:187–207
3. http://www.enterface.net
4. Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63:561–580
5. Bozkurt B (2005) New spectral methods for the analysis of source/filter characteristics of speech signals. PhD thesis, Faculté Polytechnique de Mons
6. Henrich N (2001) Etude de la source glottique en voix parlée et chantée: modélisation et estimation, mesures acoustiques et electroglottographiques, perception. PhD thesis, Université de Paris VI
7. Doval B, d'Alessandro C, Henrich N (2006) The spectrum of glottal flow models. Acta Acustica 92:1026–1046
8. Doval B, d'Alessandro C (2003) The voice source as a causal/anticausal linear filter. In: Proceedings of Voqual'03, voice quality: functions, analysis and synthesis, ISCA workshop
9. Sundberg J (1974) Articulatory interpretation of the singing formant. J Acoust Soc Am 55:838–844
10. Boite R, Bourlard H, Dutoit T, Hancq J, Leich H (2000) Traitement de la parole
11. http://www.phon.ucl.ac.uk/home/sampa/
12. Bozkurt B, Couvreur L, Dutoit T (2007) Chirp group delay analysis of speech signals. Speech Commun 49(3):159–176
13. Dubuisson T, Dutoit T (2007) Improvement of source-tract decomposition of speech using analogy with LF model for glottal source and tube model for vocal tract. In: Proceedings of models and analysis of vocal emissions for biomedical application workshop, pp 119–122
14. Edelman A, Murakami H (1995) Polynomial roots from companion matrix eigenvalues. Math Comput 64(210):763–776
15. Bozkurt B, Doval B, d'Alessandro C, Dutoit T (2005) Zeros of the Z-transform representation with application to source-filter separation in speech. IEEE Signal Process Lett 12(4):344–347
16. Fant G, Liljencrants J, Lin Q (1985) A four-parameter model of glottal flow. STL-QPSR 4:1–13
17. Fant G (1960) Acoustic theory of speech production. Mouton and Co, Netherlands
18. Vincent D, Rosec O, Chonavel T (2005) Estimation of LF glottal source parameters based on ARX model. In: Proceedings of Interspeech, Lisbonne, pp 333–336
19. Vincent D, Rosec O, Chonavel T (2007) A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling. In: Proceedings of ICASSP, Honolulu, pp 525–528
20. http://www.cycling74.com
21. http://www.puredata.org
22. d'Alessandro N, Dutoit T (2007) HandSketch bi-manual controller. In: Proceedings of NIME, pp 78–81
23. Schwarz D, Wright M (2000) Extensions and applications of the SDIF sound description interchange format. In: International computer music conference
24. d'Alessandro N, Doval B, Beux SL, Woodruff P, Fabre Y, d'Alessandro C, Dutoit T (2007) Realtime and accurate musical control of expression in singing synthesis. J Multimodal User Interfaces 1(1):31–39
25. d'Alessandro N, Dutoit T (2007) RAMCESS/HandSketch: a multi-representation framework for realtime and expressive singing synthesis. In: Proceedings of Interspeech'07, pp TuC. SS–5
26. Birkholz P, Steiner I, Breuer S (2007) Control concepts for articulatory speech synthesis. In: Proceedings of the 6th ISCA workshop on speech synthesis
27. Berndtsson G, Sundberg J (1993) The MUSSE DIG singing synthesis. In: Proceedings of the Stockholm music acoustics conference, pp 279–281
28. d'Alessandro N, Dubuisson T, Moinet A, Dutoit T (2007) Causal/anticausal decomposition for mixed-phase description of brass and bowed string sounds. In: Proceedings of international computer music conference, pp 465–468