The Speech Conductor: Gestural Control of Speech Synthesis

Christophe d'Alessandro (1), Nicolas D'Alessandro (2), Sylvain Le Beux (1), Juraj Simko (3),
Feride Çetin(4), Hannes Pirker (5)
(1) LIMSI-CNRS, Orsay, France, (2) FPMS, Mons, Belgium (3) UCD, Dublin, Ireland,
(4) Koç Univ.,Istanbul, Turkey, (5), OFAI, Vienna, Austria

Abstract

The Speech Conductor project aimed at developing a gesture interface for driving ("conducting") a speech synthesis system. Four real-time gesture controlled synthesis systems have been developed. For the first two systems, the efforts focused on high quality voice source synthesis. These "Baby Synthesizers" are based on formant synthesis and they include refined voice source components. One of them is based on an augmented LF model (including an aperiodic component), the other one is based on a Causal/Anticausal Linear Model of the voice source (CALM) also augmented with an aperiodic component. The two other systems are able to utter unrestricted speech. They are based on the MaxMBROLA and MidiMBROLA applications. All these systems are controlled by various gesture devices. Informal testing and public demonstrations showed that very natural and expressive synthetic voices can be produced in real time by some combination of input devices/synthesis system

***Index Terms*—speech synthesis, glottal flow, gesture control, expressive speech.**

I. PRESENTATION OF THE PROJECT

A. Introduction

Speech synthesis quality seems nowadays acceptable for applications like text reading or information playback. However, these reading machines lack expressivity. This is not only a matter of corpus size, computer memory or computer speed. A speech synthesizer using several times more resources than currently available will probably improve on some points (less discontinuities, more smoothness, better sound) but expression is made of real time subtle variations according to the context and to the situation. In daily life, vocal expressions of strong emotions like anger, fear or despair are rather the exception than the rule. Then a synthesis system should be able to deal with subtle continuous expressive variations rather than clear cut emotions. Fundamental questions concerning expression in speech are still unanswered, and to some point even not stated. Expressive speech synthesis is the next challenge. Expressive speech synthesis may be viewed from two sides: on the one hand is the question of expression specification (what is the suited expression in a particular situation?) and on the other hand is the question of expression realisation (how is the specified expression actually implemented). The first problem (situation analysis and expression specification) is one of the most difficult problems for research in computational linguistics because it involves deep understanding of the text and its context. Without a deep knowledge of the situation defining an adequate expression is difficult, if not impossible.
It is only the second problem that has been addressed in this workshop. Given the expressive specifications, produced and controlled in real time by a "speech conductor", given the intended expression, or an "expression score" for a given speech utterance, how to "interpret" the speech produced according to this intended expression?
The Speech Conductor project aims at developing and testing gesture interfaces for driving ("conducting") a speech or voice synthesis system. The goal is to modify speech synthesis in real time according to the gestures of the "Speech Conductor". The Speech Conductor adds expressivity to the speech flow using speech signal synthesis and modification algorithms and gesture interpretation algorithms. This multimodal project involves sounds, gestures and text.

B. Domains and challenges

The main goal of this project was to test various gesture interfaces for driving a speech synthesiser and then to study whether more "natural" expressive speech (as compared to rule-based or corpus-based approaches) could be produced. The problems addressed during the workshop were:

1. Identify the parameters of expressive speech and their relative importance. All the speech parameters are supposed to vary in expressive speech. In time domain a list of speech parameters would encompass: articulation parameters (speed of articulation, formant trajectories, articulation loci, noise bursts, etc.) phonation parameters (fundamental frequency, durations, amplitude of voicing, glottal source parameters, degree of voicing and source noise etc.). Alternatively, physical parameters (sub glottal pressure, larynx tension) or spectral domain parameters could be used.
2. Signal processing for expressive speech. Techniques for parametric modification of speech: fundamental frequency, duration, articulation, voice source.
3. Domain of variation and typical patterns for expressive speech parameters, analysis of expressive speech.

4. Gesture capturing and sensors. Many types of sensor and gesture interfaces were available. The most appropriates have been selected and tried.
5. Mapping between gestures and speech parameters. The correspondence between gestures and parametric modifications is of paramount importance. This correspondence can be more or less complex (one to many, many to one, one to one). A physiologically inspired model for intonation synthesis has been used.
6. Different types of vocal synthesis have been used. Parametric source/filter synthesis proved useful for accurately controlling voice source parameters. Diphone based concatenative speech synthesis proved useful for more unrestricted speech synthesis applications, but allowed for less fine grained controls. Of course real time implementations of the synthesis systems were needed.
7. Expression, emotion, attitude, phonostylistics. Questions and hypotheses in the domain of emotion research and phonostylistics, evaluation methodology for expressive speech synthesis have only marginally been addressed because of the short time available. For the same reason preliminary evaluation of the results obtained took place on an informal basis only.

C. Gesture Control Devices

Several devices, whose controllers and ranges are quite different, where used. At first, we used two keyboards, one Roland PC-200, with 49 keys, a Pitch Bend /Modulation Wheel and one fader. The range of the keyboard is by default between 36 and 84 but can be shifted in order to change the frequency register. The Pitch Bend/Modulation wheel sends values between 0 and 127 according to the MIDI protocol. Thus, these several controllers are respectively sending values on dedicated Note On/Off, Pitch Bend and Control Change channels.

The second keyboard was a Edirol PCR-50 which features 8 knobs and 8 faders in addition to the controls mentioned before. Similarly, in this keyboard the values are set between 0 and 127 and it sends data on several Control Change channels.

In addition to the Roland keyboard we also used an Eobody controller to have some extra knob controls in order to drive the MaxMBROLA Text-To-Speech synthesizer. This sensor interface converts any sensor raw data to MIDI protocol, but as a matter of fact we only used the inbox knobs. We were also able to use a MIDI foot controller providing ten switches in ten different banks and two expression pedals.

A P5 Glove with five flexion sensors linked to the fingers that could bend when fist clench was also employed. The sensors send data in range 0 to 63. Thanks to an Infrared sensor, the glove offers the ability to track the hand position in three spatial dimensions (x,y,z) within a continuous range roughly equal to [–500,+500].

The glove does not actually use MIDI protocol but Open Sound Control (OSC) instead. Contrary to MIDI which sets data in a serial way, under OSC the values are sent in parallel, allowing a fixed rate for every controller.

D. Overview of the work done

The work has been organized along two main lines: text-to-speech synthesis and parametric voice quality synthesis. As for text-to-speech synthesis two different configurations have been produced. For one of the systems the only parameter controlled in real time is fundamental frequency. Phonemes and durations are computed automatically by the text-to-speech engine (we used Mary (Schröder & Trouvain, 2003) for English) and then produced by the MBROLA diphone system (Dutoit & al., 1996). For the second system, syllables are triggered by the player. Then durations, fundamental frequency and intensity are controlled using the MidiMBROLA synthesis system (D'Alessandro & al. 2005).

As for parametric voice quality synthesis, coined herein the "Baby Synthesizers" also two different approaches have also been implemented. Both are based on a parametric description of the voice source. In one system, the well-known LF model (Fant & al. 1985, Fant 1995) of the glottal flow derivative has been used, and augmented with an aperiodic component. The other system is based on a spectral approach to glottal flow modelling, the Causal/Anticausal Linear Model, CALM (Doval & al. 2003). This model has also been augmented with an aperiodic component.

In the remaining of this paper, the four systems developed during the workshop will be described in more detail.

II. Real time control of an augmented LF-model.

A. The voice source model in the time domain

In the linear acoustic model of speech production, the effect of the voice source is represented by the time-varying acoustic flow passing through the glottis. When the vocal folds are regularly oscillating (voiced speech), the glottal flow can be represented using a glottal flow model, the most widely used being the Liljencrants-Fant (LF) model (Fant & al. 1985). The glottal flow is the air stream coming from the lungs through the trachea and pulsed by the glottal vibration. All the glottal flow models are pulse like, positive (except in the case of ingressive speech), quasi-periodic, continuous, and differentiable (except at closure). Acoustic radiation of speech at the mouth opening can be approximated as a derivation of the glottal flow. Therefore, the glottal flow derivative is often considered in place of the glottal flow itself. The form of the glottal flow derivative can often be recognized in the speech waveform, with additional formant ripples. The time-domain glottal flow models can be described by equivalent sets of 5 parameters (Doval & d'Alessandro, 1999):

- A_v: peak amplitude of the glottal flow, or amplitude of voicing.
- T_0: fundamental period (inverse of F_0)
- O_q: open quotient, defined as the ratio between the glottal open time and the fundamental period. This

quotient is also defining the glottal closure instant at time O_q*T_0.

- A_m: asymmetry coefficient, defined as the ratio between the flow opening time and the open time. This quotient is also defining the instant T_m of maximum of the glottal flow, relative to T_0 and O_q ($T_m = A_m*O_q*T_0$). Another equivalent parameter is the speed quotient S_q, defined as the ratio between opening and closing times, $A_m = S_q / (1 + S_q)$.
- Q_a: the return phase quotient defined as the ratio between the effective return phase duration (i.e. the duration between the glottal closure instant, and effective closure) and the closed phase duration. In case of abrupt closure $Q_a = 0$.

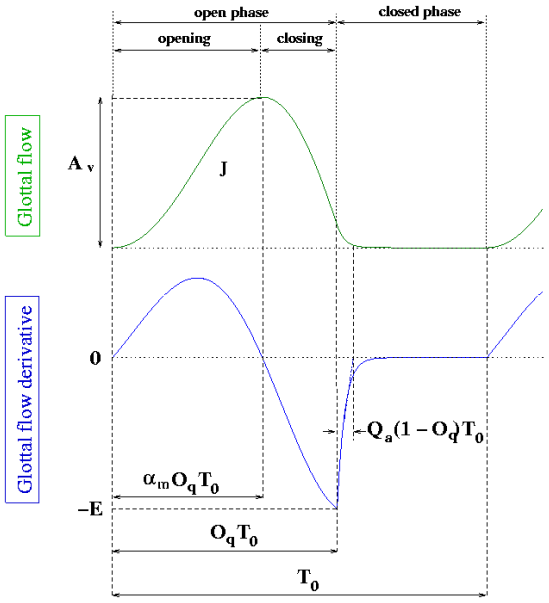


Figure 1: Time domain models of the glottal flow and glottal flow derivative (LF-model),after Henrich & al. 2002.

When considering the glottal flow derivative, the peak amplitude is generally negative, because the closing phase is generally shorter than the opening phase. So the descending slope of the glottal flow is steeper, and its derivative larger. All the time-domain parameters are equivalent for the glottal flow and its derivative except this amplitude parameter:

- E: peak amplitude of the derivative, or maximum closure speed of the glottal flow. Note that E is situated at O_q*T_0, or glottal closure instant. It is often assumed that E represents the maximum acoustic excitation of the vocal tract

E and A_v are both representing a time domain amplitude parameter. One or the other can be used for controlling amplitude, but E appears more consistently related to loudness and should probably be preferred for synthesis. The waveform and derivative waveform of the LF model are plotted in Figure 1.It must be pointed out that an aperiodic component must also be added to the periodic LF model. Two types of aperiodicities have to be considered: structural aperiodicities (jitter and shimmer) that are perturbations of the waveform periodicity and amplitude, and additive noise.

Note that compared to the LF model new parameters are added for controlling the aperiodic component. Shimmer and Jitter are perturbation of T_0 amplitude of the LF model (structural aperiodicities). Filtered white noise is also added to the source for simulating aspiration noise in the voice source. The voice source waveform is then passed in a vocal tract filter to produce vowels. The initial formant transitions have been designed to produce a voiced stop consonant close to /d/. This time-domain "baby synthesizer" based on the augmented LF model is presented in Figure 2. The circles indicate those parameters than can be controlled in real time by the gesture captors

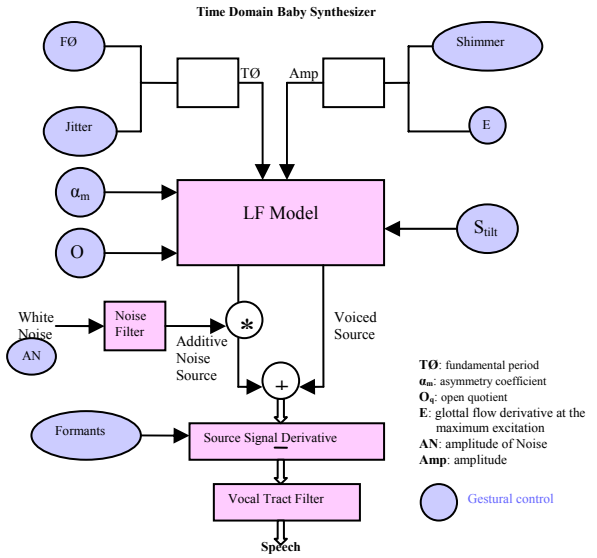


Figure 2. The time-domain "baby synthesizer" implemented in the project, LF model of the source, source aperiodicities and vocal tract filter.

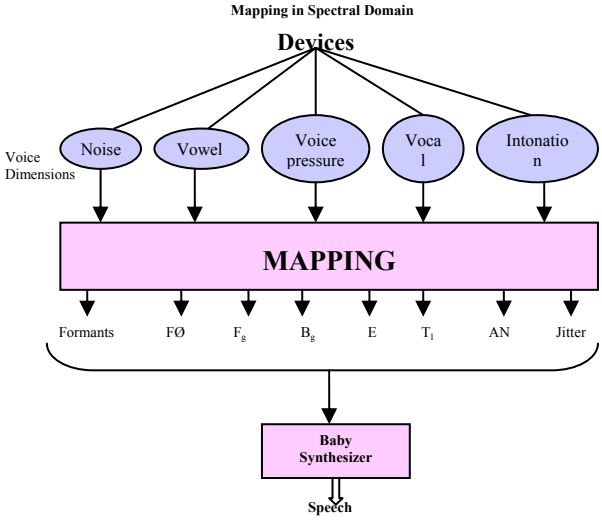
B. Mapping

There is no one-to-one correspondence between voice quality and glottal flow parameters. Their relationships are the subject of a large body of work. They can be sketched as follows (d'Alessandro, forthcoming). F_0 describes melody. A very low F_0 generally signals creaky voice and a high F_0 generally signals falsetto voice. O_q describes mainly the lax-tense dimension. O_q is close to 1 for a lax voice, and may be as low as 0.3 for very pressed or tense phonation. As A_v represents the maximum flow, it is an indication of flow voice, and it may help for analysis of the vocal effort dimension. E correlates well with the sound intensity. Q_a

correlates also with the effort dimension. When Q_a =0 the vocal cords close abruptly. Then both E the asymmetry A_m are generally high, and so is vocal effort. Conversely, large values of Q_a (0.05-0.2) give birth to a smooth glottal closure –the vocal effort is low. The asymmetry coefficient A_m has an effect on both the lax-tense dimension (asymmetry is close to 0.5 for a lax voice, and higher for a tense voice) and the vocal effort dimension (asymmetry generally increases when the vocal effort increases). Therefore some sort of mapping between raw voice source parameters and voice quality dimensions is needed.

For controlling of the baby synthesizers, voice quality dimensions are mapped onto voice source acoustic parameters. These voice quality dimensions are then controlled by the gesture captors, as explained in Figure 3.

Figure 3. Mapping in Time domain



C. Gestural control

The augmented LF model has been implemented entirely in the Pure Data environment. The implementation is based on the normalized LF model worked out in (Doval & d'Alessandro 1999).

The way controllers have been mapped to the various synthesizers was somewhat arbitrary. It must be pointed out that controllers could practically be driving any of the several synthesizers we implemented. For the augmented LF model Baby synthesizer the configuration was settled as follows:

- The Edirol MIDI keyboard was driving three voice dimensions. The keys from (from left to right) define the vocal effort, and the velocity of the pressed key was linked to the glottal pressure.
- In order to be able to have a dynamic mapping of these two dimensions we chose to have the possibility to change the parameters driving these dimensions. So that we could easily set the mid value and the span of asymmetry, open quotient and closing phase time, these parameters were each set by two knobs.
- The Pitch Bend/Modulation wheel was respectively controlling Frequency and Volume in such a way that no sound is produced the wheel is released.
- In addition to this, we used the pedal board to switch between the different presets of the vocal tract formants of different predefined vowels (a,e,i,o,u).
- Finally, one expression pedal of this pedal board was use to add noise to the signal generated.

III. Real time control of a Causal/Anticausal Linear Spectral Model

A. The voice source model in the spectral domain

Modelling the voice source in the spectral domain is interesting and useful because the spectral description of sounds is closer to auditory perception. Time-domain and frequency domain descriptions of the glottal flow are equivalent only if both the amplitude and the phase spectrum are taken into account, as it is the case in this work.

The voice source in the spectral domain can be considered as a low-pass system. It means that the energy of the voice source is mainly concentrated in low frequencies (recall that only frequencies below 3.5 kHz were used in wired phones) and is rapidly decreasing when frequency increases. The spectral slope, or spectral tilt, in the radiated speech spectrum (which is strongly related to the source derivative) is at most -6 dB/octave for high frequencies. As this slope is of +6 dB/octave at frequency 0, the overall shape of the spectrum is a broad spectral peak. This peak has a maximum, mostly similar in shape to vocal tract resonance peaks (but different in nature). This peak shall be called here the "glottal formant'". This formant is often noticeable in speech spectrograms, where it is referred at as the "voice bar", or glottal formant below the first vocal tract formant.

Spectral properties of the source can then be studied in terms of properties of this glottal formant. These properties are:

1. the position of the glottal formant (or "frequency");
2. the width of the glottal formant (or "bandwith");
3. the high frequency slope of the glottal formant, or "spectral tilt";
4. the height of the glottal formant, or "amplitude".

One can show that the frequency of the glottal formant is inversely proportional to the open quotient O_q (Doval et al. 1997). It means that the glottal formant is low for a lax voice, with a high open quotient. Conversely, a tense voice has a high glottal formant, because open quotient is low.

The glottal formant amplitude is directly proportional to the amplitude of voicing. The width of the glottal formant is linked to the asymmetry of the glottal waveform. The relation is not simple, but one can assume that a symmetric waveform (a low S_q) results is a narrower and slightly lower glottal formant. Conversely, a higher asymmetry results in a broader and slightly higher glottal formant

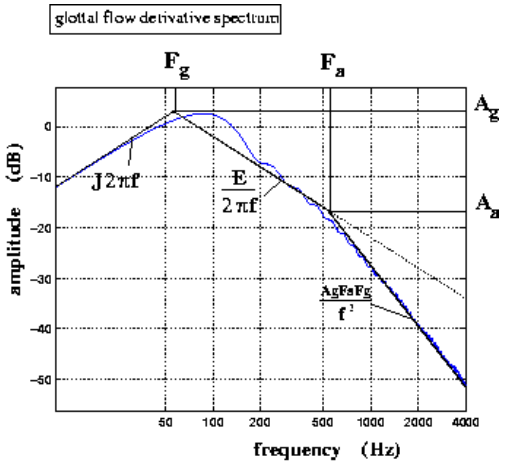


Figure 4. Glottal flow derivative spectrum (after Henrich &al. 2002)

Around a typical value of the asymmetry coefficient (2/3) and for normal values of open quotient (between 0.5 and 1), the glottal formant is located slightly below or close to the first harmonic ($H_1 = f_0$). For O_q=0.4 and A_m=0.9, for instance, it can then reach the fourth harmonic

Up to now, we have assumed an abrupt closure of the vocal folds. A smooth closure of the vocal folds is obtained by a positive Q_a in time domain. In spectral domain, the effect of a smooth closure is to increase spectral tilt. The frequency position where this additional attenuation starts is inversely proportional to Q_a. For a low Q_a, attenuation affects only high frequencies, because the corresponding point in the spectrum is high. For a high Q_a, this attenuation changes frequencies starting at a lower point in the spectrum.

In summary, the spectral envelope of glottal flow models can be considered as the gain of a low-pass filter. The spectral envelope of the derivative can then be considered as the gain of a band-pass filter. The source spectrum can be stylized by 3 linear segments with +6dB/octave, -6dB/octave and -12dB/octave (or sometimes -18dB/oct) slopes respectively. The two breakpoints in the spectrum correspond to the glottal spectral peak and the spectral tilt cut-off frequency

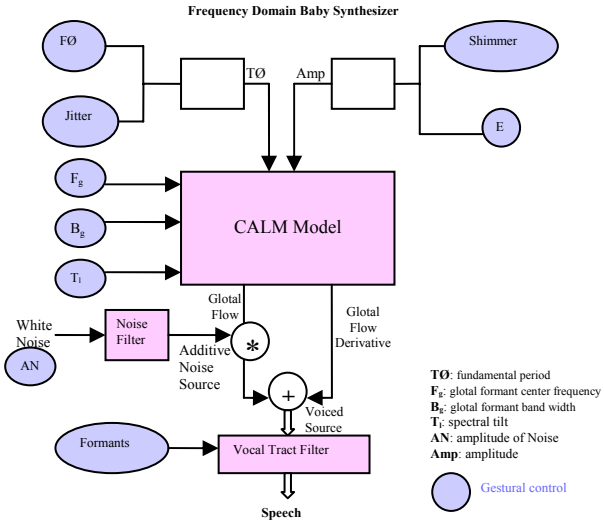
An example displaying linear stylization of the envelope of the glottal spectrum in a log representation is given in Figure 4.

For synthesis in the spectral domain, it is possible to design an all-pole filter which is comparable to e.g. the LF model. This filter is a 3rd order low-pass filter, with a pair of conjugate complex poles, and a simple real pole. The simple real pole is given directly by the spectral tilt parameter. It is mainly effective in the medium and high frequencies of the spectrum. The pair of complex-conjugate poles is used for modeling the glottal formant. If one wants to preserve the glottal pulse shape, and then the glottal flow phase spectrum, it is necessary to design an anticausal filter for this poles pair. If one wants to preserve the finite duration property of the glottal pulse, it is necessary to truncate the impulse response of the filter. The spectral model is then a Causal (spectral tilt) Anti-causal (glottal formant) Linear filter Model (CALM, see Doval & al. 2003). This model is computed by filtering a pulse train by a causal second order system, computed according to the frequency and bandwidth of the glottal formant, whose response is reversed in time to obtain an anti-causal response. Spectral tilt is introduced by filtering this anti-causal response by the spectral tilt component of the model. The waveform is then normalized in order to control accurately the intensity parameter E.

An aperiodic component is added to this model, including jitter, shimmer and additive filtered white noise. The additive noise is also modulated by the glottal waveform.

Then the voice source signal is passed through a vocal tract formant filter to produce various vowels. Figure 6 presents an overview of the spectral "Baby synthesizer.

Figure 6. CALM Model



B. Mapping

This global spectral description of the source spectrum shows that the two main effects of the source are affecting the two sides of the frequency axis. The low-frequency effect of the source, related to the lax-tense dimension is often described in terms of the first harmonic amplitudes H_1 and H_2 or in terms of the low frequency spectral envelope. A pressed voice has a higher H_2 compared to H, and conversely a lax voice has a higher H_1 compared to H. The effort dimension is often described in terms of spectral tilt. A louder voice has a lower spectral tilt, and spectral tilt increases when loudness is lowering.

Then the vocal effort dimension is mainly mapped onto the spectral tilt and glottal formant bandwidth parameters (asymmetry), although the voice pressure dimension depends mostly on the glottal formant centre frequency, associated to open quotient.

Other parameters of interest are structural aperiodicities (jitter and shimmer) and additive noise.

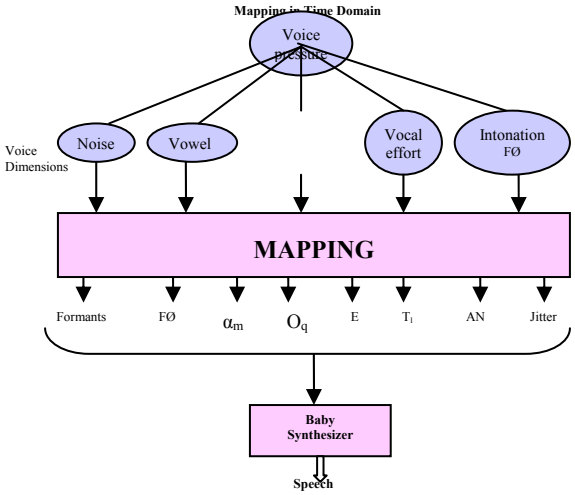


Figure 7. Mapping in Spectral domain

C. *Gestural control of the spectral Baby Synthesizer*

For this synthesizer, a P5 data glove is used. This input device allows driving 8 continuous variable parameters at once: 3 spatial position x, y, z associated with the movement of the glove relative to a fixed device on the table and 5 parameters associated with bending of the five fingers. Several keys on the computer keyboard are controlling vowels. The glove was driving the spectral-domain glottal source model. Only the two horizontal spatial dimensions (x,z) were used as follows: the x variable was linked to intensity E and the z variable was linked to fundamental frequency. All the fingers but the little finger were used to control respectively (beginning from the thumb) noise ratio, Open Quotient, Spectral Tilt and Asymmetry. This mapping is most reliable and effective (compared to the keyboard used in the first experiment). Only a short training phase was sufficient to obtain very natural voice source variations. The computer keyboard was used for changing values of the formant filters for synthesizing different vowels, and then basic vocal tract articulations.

IV. Real time control of F0 in a Text-To-Speech System using MaxMbrola

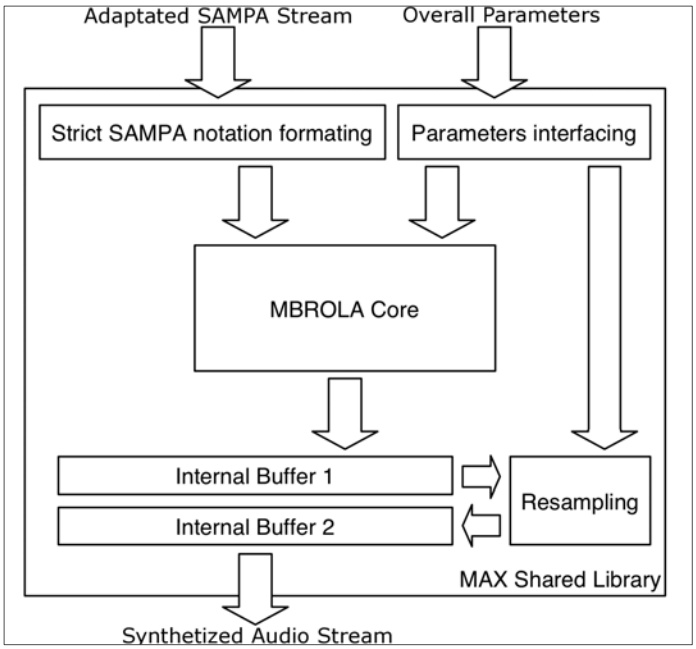
A. *Max/MSP Graphical Programming Environnement*

The Max graphical development environment and its MSP audio processing library (Zicarelli & al., 2004) are widely used the computer music community. This software is a powerful tool in many fields of electronic music like real-time sound processing, control mapping, composition, enhancement of performance abilities etc. It is a rare example of an intuitive interface (design of personalized modules by the building of graphs of simple functions, called *objects*) and a high level of flexibility (functions accepting and modifying numbers, symbols, audio and video stream, etc) at the same time. The capabilities of that software increase every day due to the help of an active developer community providing new *external* objects (or *externals*).

B. *MaxMBROLA~ external object: MBROLA inside Max/MSP*

This section explains how the MBROLA technology has been integrated inside the Max/MSP environment (D'Alessandro & al. 2005). Max/MSP objects work as small servers. They are initialized when they are imported into the workspace. They contain a set of dedicated functions (methods) which are activated when the object receives particular messages. These messages can be simple numbers, symbols or complex messages with a header and arguments. Considering that real-time request-based protocol of communication between objects, a Max/MSP external object containing the MBROLA algorithm has been developed and a particular set of messages (header and arguments) has been formalized to communicate with the synthesizer.

Figure 8. Internal structure of the MaxMBROLA~ external object (after D'Alessandro & al. 2005).



As shown in Figure 8, we can separate the possible requests in two main channels. On one side, there is parameter modification, which influences the internal state of the synthesizer. On the other side, there is the phonetic/prosodic stream, which generates speech instantaneously.

C. *Available actions of the object*

1) *Internal state modifications*

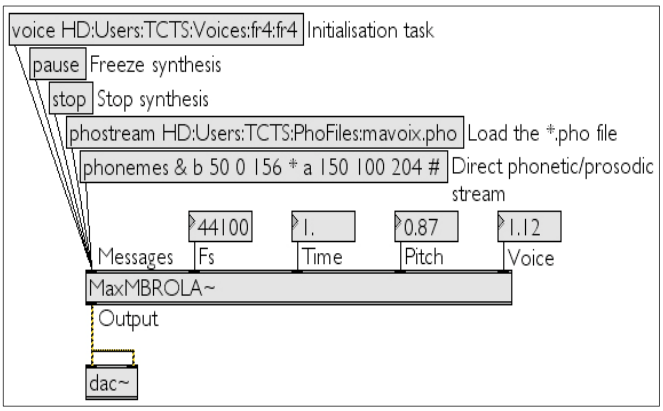
Specific modifications of the internal state of the MBROLA synthesizer can be applied with Max/MSP requests. Here follows a description of the supported actions. The labels are used to name inlets (from left to right: *Messages*, *Fs*, *Time*, *Pitch* and *Voice)* and examples of the supported messages are illustrated on Figure 9.

The synthesizer always starts with the initialization task (*Messages* inlet). Thit function starts the MBROLA engine loads the requested diphone database and set all the internal

parameters to their default values. All the existing MBROLA databases are compatible with this application.

The stream provided by the external can be frozen (*Messages* inlet). It means that the phonetic/prosodic content stays in memory but the MBROLA engine stops the synthesis task.

Figure 9. Supported messages of the MaxMBROLA~ external object.



The MBROLA engine can also be stopped (*Messages* inlet). That function flushes the phonetic/prosodic content, stops the synthesis process and sets all the internal parameters to their default values. The diphone database remains loaded.

Fs inlet receives a floating point number. It controls the output sampling rate. Indeed, the original sampling rate depends on the database (16000Hz or 22050Hz). Linear interpolation is performed allowing the use of that external object with all possible sampling rates.

The inlets *Time*, *Pitch* and *Voice* each receive a floating point number. These values are respectively the time ratio (deviation of the reference speed of speech), the pitch ratio (deviation of the reference fundamental frequency of speech) and voice ratio (compression/dilation ratio of the spectrum width). For each inlet, 1.0 is the default value. The object doesn't transmit values lower than 0.01 (means "100 time lower than the default value").

2) Phonetic/prosodic stream processing

The requests for generating speech in the Max environment are described. All the following messages are sent into the *Messages* inlet.

A loading request allows to use a standard *.pho file (which include the list of phonemes to be produced and the target prosody) to perform synthesis. Examples are available together with MBROLA voices and complete explanations about standard SAMPA (Speech Assessment Methods Phonetic Alphabet). SAMPA is a machine-readable phonetic alphabet used in many speech synthesizers. (Cf. the SAMPA-page http://www.phon.ucl.ac.uk/home/sampa/home.htm).

We developed a function that directly accepts SAMPA streams inside Max messages to provide user control to interactive speech production. The standard SAMPA notation has been modified to fit to the Max message structure. For example, the following stream:

phonemes & b 50 0 156 * a 150 100 204 #

begins by initializing the synthesizer, then produces a syllable /ba/ of 200 (50 + 150) milliseconds with a fundamental frequency increasing from 156Hz to 204Hz (two pitch points). Finally, it flushes the phoneme buffer.

D. Adding Text-to-Phoneme capabilities to MaxMBROLA

MaxMBROLA requires a phonemic specification as input just like it is used in mbrola .pho files, i.e. a transcription in SAMPA with optional information on duration and pitch. MaxMBROLA, just as mbrola, is not intended to be a fully fledged text-to-speech system. Anyway, it is obviously advantageous to combine it more directly with some kind of text-to-phoneme preprocessing in order to increase the flexibility of the system.

It was thus decided to use the text-to-phoneme capabilities provided by the TTS-system Mary (Schröder & Trouvain, 2003).

Mary is a Text-To-Speech system available for German and English. One of its attractive properties is that it offers full access to the results of intermediate processing steps. It provides an XML representation that contains not only the phonemes, their durations and pitch, but also a straightforward encoding of the full prosodic hierarchy which comprises phrases, words and syllables.

As there are applications of MaxMBROLA where the speech is to be synthesized syllable-wise, the latter information is most valuable.

A collection of simple Perl-scripts for parsing and converting Mary-XML format as well as standard mbrola .pho files to the input format required by MaxMBROLA was produced.

Max/MSP provides a "shell"-object which allows the execution of shell-commands, including piping, within a patch. This made the smooth integration of the text-to-phoneme processing rather straightforward.

As Mary is implemented as server-client architecture, as a special treat Mary was currently not installed locally but was accessed via Internet from within Max/MSP.

E. Gestural control of the Text-to-Speech system

Only one parameter, namely fundamental frequency (F0), was controlled by the glove in the MaxMbrola + mary text-to-Speech system. The phoneme stream and segment durations were computed by the TTS system. A flat pitch MBROLA signal was computed according to this data. Then F0 movements were computed by a PSOLA post-processing module receiving the flat MBROLA synthesized speech as input. F0 was modulated in real time, according to the distance between the glove and a fixed device on the table. This very simple control scheme was very effective. Very realistic and expressive prosodic variations were produced almost immediately because controlling F0 this way proved very intuitive.

V. Real time control of F0, durations and intensity in a syllable based Speech synthesis System using MidiMbrola

A. MIDI-MBROLA: The First MaxMBROLA-based MIDI Instrument

A Max/MSP musical instrument, called MIDI-MBROLA, has also been developed around the MaxMBROLA external object (D'Alessandro & al. 2005). This tool has a full MIDI compatible interface. MIDI *control changes* are used to modify the internal parameters of the MBROLA synthesizer. *Events* from a MIDI keyboard are used to compute the prosody, which is mixed with the phonetic content at the time of performance. As a standard module of the Max/MSP environment, the MIDI-MBROLA digital instrument automatically allows polyphony. Indeed, many voices can readily be synthesized simultaneously because the MBROLA synthesis doesn't utilize many CPU resources. It can also be compiled as a standalone application or as a VST instrument ("Virtual Studio Technology", a digital effect standard developed by Steinberg) instrument. That tool is publicly available.

B. Gestural control of MIDI-MBROLA

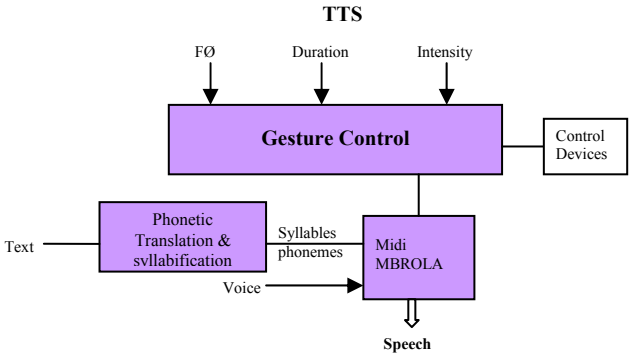
The MIDI-MBROLA instrument has been linked to the Roland keyboard and the three knobs of the Eobody Controller. The input text consisted of a syllabic sliced phonetic transcription of the speech utterance. Syllables were triggered by the keyboard. F0 was modulated by the keyboard and pitch-bend. Note that the keyboard has been divided in 1/3 of semitone between to adjacent keys. The Pitch Bend allowed for even smaller pitch excursions. The three knobs were controlling the overall speed, the mid-pitch and the vowel length. But it should be noticed that only the pitch control was effectively driving a parameter in real time whereas the three others were only sampled at syllables frequency (his means that once triggered a syllable was played with a given speed, without variation within the syllable). The configuration used is showed in Figure 9. With this configuration, the output speech had a singing character which sounded rather unnatural for speech. This was because the pitch variations were limited by the discrete nature of the keyboard.

VI. Fujisaki Intonation Modelling

Another strand of development dealt with the implementation of the Fujisaki model of intonation (Fujisak &Hirose, 1984) in the Pure Data environment. This model aims to take physical and physiological processes involved in the production of F0 into account. The main idea is to model the intonation contour by superimposing the results of two different processes. On the one hand there is the phrase component that models the phenomenon of slowly declining global pitch baseline throughout a prosodic phrase. The accent component, on the other hand, is responsible for modeling the local excursions in the F0 contour used for marking pitch accents.

Fujisaki's model has proved its descriptive adequacy in capturing F0 contours for a variety of different languages.
As the input parameters of the model that have to be dynamically controlled can be basically reduced to the triggering of phrase commands, accent commands and their respective amplitudes, it seemed worthwhile to investigate its applicability in a real-time system.
An implementation of the Fujisaki in PureData was produced. In a first experiment the parameters where controlled by a MIDI-keyboard, where attack, release and velocity map quite straightforwardly to the timing and the amplitude of both accent- and phrase commands.

Figure 10. TTS Control Model



VII. Discussion and conclusion

A. Summary of software produced.

Four different software projects have been produced during eNTERFACE:

1. the time-domain Baby Synthesizers. A LF model based vowel formant synthesizer, written in Pure Data, and mainly tested with keyboard, joystick and pedal-board real-time interfaces.
2. the spectral domain Baby synthesizer. A CALM model based vowel formant synthesizer, written in Max/MSP, and mainly tested with a digital glove real-time interface.
3. the Mary TTS in English with real-time intonation control, using a digital glove.
4. the MIDI-MBROLA speech synthesizer in French with a real-time control of intonation, duration and intensity using a keyboard with pitch bend.

B. Comparing patch programming Environments

Baby Synthesizers were developed using the real-time graphical environments Pure Data (PD) and Max/MSP. PD is an Open Source platform developed and maintained by Miller Puckette and includes code written by wide community of programmers. Max/MSP is commercial software developed by Cycling'74 company.
During this process we also tested some limits of these closely related platforms, and learnt lessons which we share.
Graphical environment
Being a commercial product, Max/MSP environment is better designed and user friendlier. However, simpler PD user

interface wasn't causing any problems in development process.

Stability

No stability issues with Max platform were encountered during the development. On the other hand, Pure Data programmers experienced several challenging problems, when some objects kept changing their behavior, disappearing and reappearing randomly. In general, stability issues were less serious for MacOS then for Windows platform; even system reboot didn't always help…

Richness

PD proved to be slightly more flexible when it came to coding more complex mathematical functions on sound wave in real time. Unlike Max/MSP, it allows a wide variety of mathematical operations to be performed in real-time directly on the sound signal with one very simple universal object. Similar operations had to be coded in C, compiled and imported to MAX/MSP.

Despite of the limitations mentioned above, both of these closely related environments proved to be suitable for sound processing applications of the kind we were developing.

C. Towards expressivity evaluation

Up to now no formal evaluation of the different variants of synthesizers has been performed. As a matter of fact, the evaluation of the "quality" of a speech synthesis system is not a trivial task in general, and is even more complicated when it comes to the evaluation of expressivity.

Usually synthesis systems are evaluated in terms of intelligibility and "naturalness". For the former there exist a number of established tests (Gibbon et al. 1997). Typically samples of isolated syllables or nonsense words are presented and it is possible to perform a quantitative evaluation of correctly perceived samples. When evaluating the "naturalness" of synthesized speech, an objective measure is less straightforward. In the simplest case, a comparison between two systems or between two variants of a system by forced preference choice can be performed. Another method is the rating of the "adequacy" of a synthesized sample for a given context. But again it is difficult to impossible to come up with an objective independent evaluation.

In the field of the synthesis of expressive speech, the predominant evaluation method is to synthesize sentences with neutral meaning and encode a small set of "basic" emotions (typically joy, fear, anger, surprise, sadness). Subjects are then asked to identify the emotional category.

A competing evaluation model is to use more subtle expressive categories: use test sentences with non-neutral semantics, and let again rate the adequacy of the sample for a given context.

In the context of the Speech-Conductor project it was only possible to perform an informal comparison of the two synthesizers that implemented glottal source models. At the current state the CALM based model gives much better "impression" than the time-domain model. On the other hand there are still a number of slight differences in the actual implementation of these two models; e.g. the differences in the modeling of jitter and shimmer or the automatic superimposing of micro-prosodic variations, that have a strong impact on the perceived "quality" of the models.

A more interesting evaluation would be a rating test for the recognizably of perceptual voice quality measures such as laxness/tenseness, vocal effort etc. Though this would be probably a promising method of evaluating the current state, it is not easy to perform, as it would rely on the availability of independent "expert" listeners with a certain amount of phonetic experience.

In this context it would thus be interesting to further investigate whether it is possible to get reliable ratings on voice quality factors from so called "naive listeners".

For the MaxMBROLA system different evaluation methods have to be taken into account, as this is basically a classical diphone-synthesis system which allows for the real-time control of prosodic features, most prominently pitch. Thus the evaluation methods used for "normal" concatenative synthesis systems could easily applied. One of the peculiarities of this system is that inevitable the virtuosity of the person "conducting" the synthesizer is a strong factor in the quality of the output.

A straightforward evaluation would be a rating test of different input devices (e.g. Data Glove vs. Keyboard), but apart from the "human factor", currently still too many differences in the underlying synthesis scenarios exist to allow a real comparison.

D. Conclusion

Devices:

The glove performed much better than the keyboard or joysticks for controlling intonation and expressivity. However, the tested glove model had some performance limitations (it proved too slow for real time). However, the glove wasn't tested for its capacity to reproduce the intended gesture precisely and reliably.

Keyboard on the contrary allows for exact reproducibility of gestures. When combined with TTS synthesizers the produced speech had somewhat singing quality, as pitch changes are directly linked to syllable onsets.

Synthesizers:

In general, voice source models produced much more expressive vocal utterances than TTS models. For TTS, better results were reached when speech was generated using pre-computed segment durations and intensity and we only controlled F_0. So, surprisingly, less control can in some situations yield better results. In any case, it's clear that to add a real expressivity, flexible control of all of the voice source parameters is needed.

To our best knowledge, this project was the first attempt to implement real-time system of gestural control of expressive speech. The results proved really encouraging, and opened a new avenue for expressive speech synthesis research.

ACKNOWLEDGEMENTS

Hannes Pirker states that his research is carried out within the Network of Excellence Humaine (Contract No. 507422) that is funded by the European Union's Sixth Framework Programme with support from the Austrian Funds for Research and Technology Promotion for Industry (FFF 808818/2970 KA/SA). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained herein.

REFERENCES

(Bozkurt et al., 2005) B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech" IEEE Signal Processing Letters, Vol. 12, No. 4, April 2005, p 344-347

(d'Alessandro, 2005) C.d'Alessandro, "*Voice source parameters and prosodic" analysis,* in Methods in Experimental prosody research, Mouton de Gruyter (in press)

(d'Alessandro & Doval, 2003) C. d'Alessandro, B. Doval, "Voice quality modification for emotional speech synthesis", Proc. of Eurospeech 2003, Genève, Suisse, pp. 1653-1656

(D'Alessandro et al., 2005)N. D`Alessandro, B. Bozkurt, T. Dutoit, R. Sebbe, 2005, "MaxMBROLA: A Max/MSP MBROLA-Based Tool for Real-Time Voice Synthesis", Proceedings of the EUSIPCO'05 Conference, September 4-8, 2005, Antalya (Turkey).

(Doval & d'Alessandro, 1999) B. Doval, C. d'Alessandro, 1999. *The spectrum of glottal flow models.* Notes et Documents LIMSI 99-07, 22p.

(Doval & d'Alessandro, 1997) B.Doval and C.d'Alessandro. *Spectral correlates of glottal waveform models: an analytic study.* In International Conference on Acoustics, Speech and Signal Processing, ICASSP 97, pages 446--452, Munich, avril 1997. Institute of Electronics and Electrical Engineers

(Doval et al., 2003) B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," in *Proc. ISCA ITRW VOQUAL 2003*, Geneva, Switzerland, Aug. 2003, pp. 15–19

(Dutoit et al., 1996) T. Dutoit, V. Pagel, N. Pierret, F. Bataille and O. van der Vrecken, ``The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes" Proc ICSLP, Philadelphia, pp. 1393-1396, 1996.

(Fant et al., 1985) Fant G., Liljencrants J. and Lin Q. (1985) "A four-parameter model of glottal flow". STL-QPSR 4, pp. 1-13.

(Fant, 1995) G. Fant, "The LF-model revisited. Transformation and frequency domain analysis," *Speech Trans. Lab. Quarterly .Rep., Royal Inst. of Tech. Stockholm*, vol. 2-3, pp. 121-156, 1995.

(Fels, S. 1994) Fels, "Glove talk II: Mapping hand gestures to speech using neural networks," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 1994.

(Dutoit, 1997) Dutoit T. An Introduction to Text-To-Speech Synthesis. Kluwer Academic Publishers, 1997.

(Fujisaki & Hirose, 1984) H. Fujisaki, K. Hirose Analysis of voice fundamental frequency contours for declarative sentences of Japanese Journal of Acoustic Society. Jpn. (E) 5, 4. 1984

(Gibbon et al., 1997) Gibbon, D., Moore, R. & Winsky, R. (Eds) Eagles handbook of Standards and Resources for Spoken Language Systems (1997) Mouton de Gruyter

(Henrich et al. 2002) N. Henrich, C. d'Alessandro, B. doval. `` Glottal flow models: waveforms, spectra and physical measurements". Proc. Forum Acusticum 2002, Séville 2002

(MIDI, 1983) "MIDI musical instrument digital interface specification 1.0," Int. MIDI Assoc., North Hollywood, CA, 1983.

(Schröder, 2004) M. Schröder "Speech and emotion research", Phonus, Nr 7, june 2004, ISSN 0949-1791, Saarbrücken

(Schröder & Trouvain, 2003) M. Schröder & J. Trouvain (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology, 6, pp. 365-377.

(Zicarelli et al., 2004b) Zicarelli, G. Taylor, J. K. Clayton, J. and R. Dudas, MSP 4.3 Reference Manual. and Max 4.3 Reference Manual. Cycling'74/Ircam, 1990-2004.

(Wanderley & Depalle; 2004) M. Wanderley and P. Depalle, "Gestural Control of Sound Synthesis", Proc. of the IEEE, 92, 2004, p. 632-644.

http://mary.dfki.de

http://tcts.fpms.ac.be/synthesis/maxmbrola/

http://www.disc2.dk/tools/SGsurvey.html