A short review of research on voice transformations at IRCAM

P. Lanchantin, S. Farner, C. Veaux, G. Degottex, A. Roebel, X. Rodet
IRCAM
1, place Igor Stravinsky
75004 Paris, France
lanchant@ircam.fr

Abstract

IRCAM has a long experience in analysis, synthesis and transformation of voice. Natural voice transformations are of great interest for many applications and can be combine with text-to-speech system, leading to a powerful creation tool. We present research conducted at IRCAM on voice transformations for the last few years. Transformations can be achieved in a global way by modifying pitch, spectral envelope, durations etc. While it sacrifices the possibility to attain a specific target voice, the approach allows the production of new voices of a high degree of naturalness with different gender and age, modified vocal quality, or another speech style. These transformations can be applied in realtime using ircamTools TRAX.Transformation can also be done in a more specific way in order to transform a voice towards the voice of a target speaker. Finally, we present some recent research on the transformation of expressivity.

Keywords: Speaker transformation, Speaker conversion, Dynamic Model Selection, transformation of expressivity, Separation of Vocal-tract and Liljencrants-Fant model plus Noise method.

1. Introduction

Founded by Pierre Boulez in 1977, IRCAM, the Institute for Research and Coordination Acoustic / Music, is one of the world's largest public research centers dedicated to both musical expression and scientific research. It is a unique location where artistic sensibilities collide with scientific and technological innovation. It has extensive experience in analysis, transformation and synthesis of sounds and in particular of speech dating back to its beginnings [1, 2, 3, 4], and has continued until today. For the last years, research in speech was mostly oriented towards voice transformations and Text-to-Speech synthesis (TTS). For instance, IRCAM develops software SUPERVP that includes treatments specifically designed for speech [5, 6, 7, 8], the software DIPHONE STUDIO [9], which use the concept of acoustic units, PSOLA and SINOLA [10]. It recently proposed the ircamTools commercial plug-in TRAX which offers a novel approach to voice synthesis through gender and age sound transformations. IRCAM is also developing TTS systems by unit selection [12, 13, 14] and HMM-based speech synthesis [15, 16]. Other studies related to speech include the modeling of the prosody [17] and the independent modeling of glottal source and vocal tract which allows to treat them separately.

Voice transformations, TTS and their combined use have a great potential for artistic creations. Regarding the transformation of voice, IRCAM has worked on French films such as "Farinelli" by G. Corbiau, "Vatel" by R. Joffé, "Vercingétorix" by J. Dorfmann, "Tirésia" by B. Bonello, "Les Amours d'Astrée et de Céladon" by E. Rohmer, but also for "Jeu d'enfants" by Y. Samuell, "The Last Dragon" by M. Schultz. In the film industry, transformations can be useful to convert the actor's voice into another that is more suitable for the role, it can allow the use of one actor to achieve several voices in dubbing or animation movies, modification of accentuation of recorded speech, or creation of animal voices, to mention a few applications. In the music field, IRCAM has created a synthetic voice for the opera "The Mask of Orpheus" by H. Birtwistle using the software CHANT [18]. A real-time synthesis of spoken and sung choruses [19] has been developed in the Max software platform used particularly in the opera "K" by P. Manoury at Opera Bastille. IRCAM has also worked and still works on several projects including Multimedia voice. For instance, in "Les Variations Darwin", the stage director J. F. Peyret worked with IRCAM on an order of the Theatre National de Chaillot. His project used the automatic generation of text and speech processing in real time, with music by A. Markeas. Similarly, IRCAM has worked with director E. Genovese on the transformation of voice actors from the "Comédie francaise" and sound environments for the staging of "le privilege des chemins" and lastly on "Speaking" from J. Harvey.

In this paper, we will focus on voice transformations. However, for the interested reader, we note that recent research on HMM-based speech synthesis [15, 16] combined

Part of the research presented in this paper was supported by FEDER Angelstudio : Générateur d'Avatars personnalisés ; 2009-2011

with advanced modeling of the prosody [17] are promising on a compositional perspective and will definitely be of interest for composers in the next years.

The paper is organized as follows, in Section 2, we introduce the basic tools that are used for transformations, in Section 3 we present the transformations of type and nature. In Section 4, we introduce the transformation towards a target voice. In Section 5, we present transformation of expressivity. Finally we conclude and give further directions in Section 6.

2. Signal Transformation

Acoustically, the vocal organ consists of the vocal tract (mouth and nose cavities) as a resonance chamber and the larynx (the vocal folds (glottis), the false vocal folds and epiglottis) as the principal sound producing mechanism, thus called the voice source. A model relating the physics of the voice and the emitted signal is necessary for changing the speech signal. A number of signal-centered methods have been developed, the most successful ones probably being the PSOLA method [20], harmonic plus noise methods (HNM) [21], STRAIGHT [22] and the phase vocoder [7, 8, 23]. While all these methods can perform transposition and time stretching of the signal, only the latter two principles allow finer modification of the signal in the frequency domain. SUPERVP - our improved version of the phase vocoder is under continuous development for musical applications at IRCAM, and serves as the engine of AUDIOSCULPT, a powerful graphically interactive software for music modification [25]. However the optimal model is probably one than separates the glottal source (as far as possible) from the vocal tract. In [26], we recently proposed a new model in which the glottal source is separated into two components: a deterministic glottal waveform Liljencrants-Fant model and a modulated Gaussian noise.

We first present the phase vocoder and the improvements that have been made in 2.1, then we introduce the glottal source modeling in 2.2 which will be the basic tools used for the different transformations described in the following of this paper.

2.1. The phase vocoder and improvements

The basic phase vocoder [27] is roughly a series of band filters, in practice implemented as successive Short-Time Fourier Transforms (STFTs), that reduce the signal into amplitudes and phases in a uniform time-frequency grid. Combined with resampling and changing of the time step between analysis and synthesis, this method allows for high-fidelity time stretching and pitch transposition as well as modification of the amplitude of each point in the grid, and thus an enormous potential for transformations.

A well-known artifact of the phase vocoder is the introduction of "phasiness", in particular for speech, the result sounding strangely reverberant or with a lack of presence of the speaker. Improvements added to our implementation of the phase vocoder and constituting SUPERVP are: detection and processing of transients [24], waveform preservation for single-source processing [23], robust spectral-envelope estimation [5], and dynamic voicing control based on spectral-peak classification [28].

2.2. Glottal source model

Recently, we proposed an new glottal source and vocal-tract separation method called Separation of Vocal-tract and Liljencrants-Fant model plus Noise (SVLN [26]). In this method, the glottal excitation is separated into two additive components: a deterministic glottal waveform modeled by the LF model [29] and a noise component modeled by a Gaussian noise. The parametrization by only two parameters - the shape parameter Rd and the noise level σ_g - allows an intuitive control of the voice quality. Rd characterizes the slope of the glottal spectrum and can be seen as a measure of the relaxed or tensed quality of the glottis. For instance, if this slope drops sharply, this will be reflected perceptually as a relaxed voice. On the contrary, if the slope drops slowly and the glottal spectrum has a lot of treble, this will result in a rather aggressive voice and therefore perceived as a tensed voice. An estimate of the LF model [26, 30] is used to extract Rd and σ_g parameters and the VTF is estimated by taking the estimate of the glottal source into account. The VTF parameters are thus independent of the excitation parameters and the glottal source may be changed, keeping the VTF untouched which can be of interest for voice transformations in expressive speech synthesis. The speech model is the following: the signal is assumed to be stationary in a short analysis window ($\approx$ 3 periods in voiced parts, $5ms$ in unvoiced parts). In the frequency domain, the voice production model of an observed speech spectrum $S(\omega)$ is (see Fig. 1):

$$S(\omega) = (H^{f_0}(\omega) \cdot G^{Rd}(\omega) + N^{\sigma_g}(\omega)) \cdot C^{\bar{c}}(\omega) \cdot L(\omega) \quad (1)$$

H^{f_0} is a harmonic structure with fundamental frequency f_0. G^{Rd} is the deterministic excitation, i.e. an LF model [29]. This model is defined by: the fundamental period $1/f_0$, 3 shape parameters and the gain of the excitation E_e. To simplify the LF control, the parameter space is limited to a meaningful curve and a position defined by the value of Rd [29]. N^{σ_g} is a white Gaussian noise with standard deviation σ_g. C is the response of the VTF, a minimum-phase filter parametrized by cepstral coefficients $\bar{c}$ on a mel scale. To avoid a dependency between the gains E_e and σ_g on one hand and the VTF mean amplitude on the other hand, a constraint is necessary. $G^{Rd}(\omega)$ is normalized by $G^{Rd}(0)$ and E_e is therefore unnecessary. Finally, L is the lips radiation. This filter is assumed to be the time derivative ($L(\omega) = j\omega$). The estimation method for each of the parameters can be found in [26] This method and in particular the estimation of the Rd parameter but also the *Glottal*

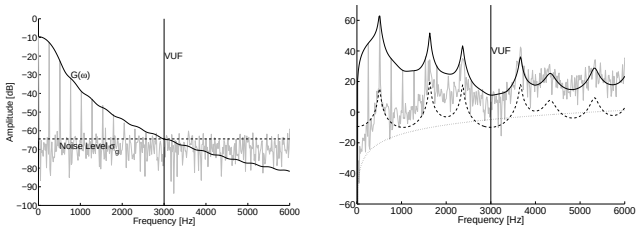

Figure 1. The mixed excitation model on the left: G^{Rd} (solid line) and the noise level σ_g (dashed line). The speech model on the right: the VTF (dashed line); L (dotted line); the spectrum of one speech period (solid line). Both plots show in gray the source $(H \cdot G + N)$ or the speech spectrum $S(\omega)$ respectively.

Closure Instant (GCI) detection algorithm which is deduced from this glottal model [26, 31], allows an intuitive control of the voice quality which is of great interest for expressive speech transformations or to quickly synthesize different speaker personalities with various voice qualities from the same voice [32]. Recently Rd estimation and GCI detection have been implemented into SUPERVP which provides a wide range of voice transformations as we will see in the following.

2.3. Basic signal analysis and transformations

We now describe some of the analysis and the transformations which will be used in the following of the paper. First, the *fundamental frequency* f_0 is an important component in the transformations, as is a robust decision of whether the signal is voiced or not [33]. Another important property of the speaking voice is the fact that the harmonics in voiced segments of the signal are masked by noise above a certain frequency, which may vary from below f_0 to half of the sampling rate depending of the voice and the phonatory setting applied. This *voiced/unvoiced frequency* will be denoted by VUF in the following. A robust estimation of the *spectral envelope* is obtained by the cepstrally based *true-envelope estimator* [34]. Compared to LPC-based methods, it has the advantages of not being biased for harmonic signals and that its order may be adapted automatically to the local f_0. The fact that the true-envelope truly follows the spectral peaks of the signal, equips us with the control necessary for detailed filtering in time and frequency depending on the time-frequency characteristics of the signal. As long as the time variation is done with care to avoid audible discontinuities, the results keeps a high degree of naturalness. Finally, as we described in the Section 2.2 the Rd parameter and the detection of GCI is an other important component in the transformations [32].

When it comes to the basic transformations, by (pitch) *transposition*, we mean changing the local f_0 of the signal by a certain factor while conserving the spectral envelope (as far as possible). The transposition of the spectral envelope is an independent parameter although both are done in the same operation. *Time-frequency filtering* has already been mentioned, and an other basic transformation is *time stretching*, which does not touch the frequency dimension. Finally, voice quality (breathy, harsh voice) can be transformed by modifying the glottal source characteristics via the Rd parameter, and GCI marks allows to introduce jitter (e.g. creaky voice).

3. Transformation of type and nature

Transformation of the voice of a given person (*source voice*) to that of another person (*target voice*), referred to as speaker conversion, has been the subject of persistent efforts in many research labs, including IRCAM as we will see in Section 4. However, it is sometimes not necessary to reach a specific voice target. In this way, an alternative approach, which has been adopted by S. Farner in [35], rather than trying to attain a specific target voice, favors the quality of the modified sound by controlling the transformation conditions. He showed that it is nevertheless possible to change the identity of the source voice by changing its apparent size, gender and age, or making the voice breathy, softer, rougher or less happy, or even reducing it to a whisper. We first present in Section 3.1 what distinguishes voices according to voice physiology and phonatory settings. In Section 3.2 we present the corpuses which were recorded to set up the different transformations presented in Section 3.3 which were implemented into TRAX, a transformation plugin presented in Section 3.4.

3.1. Differences between voices

Apart from variation in natural pitch range, different voices are distinguished and recognized by their *timbre* that depends on the physiology of the voice and the phonatory settings. The term timbre is often defined as the quality of a sound other than the pitch, duration, and loudness. For the voice we often use the term voice quality for grouping timbre-related qualities like dark, bright, soft, rich, noisy, pure, rough etc.

3.1.1. Voice physiology

The specific configuration of the voice organ, such as the length and shape of the vocal tract and the vocal folds, varies from person to person and gives them their individual pitch range and timbre. Nevertheless, there are general differences depending on the gender and age of the person [36, 37, 38], although it might be difficult in some cases to guess the gender and age merely from the person's voice. The most important differences are the natural vibration frequency range of the vocal folds (perceived as pitch and measured as f_0), the spectral distribution of the glottal source (for instance measured as spectral tilt), and the shape of the vocal tract (specific resonances and anti-resonances called formants and anti-formants).

Iseli et al. [38] have reported pitch means and ranges for male and female voices of ages ranging from 8 to 39 years: about 250 Hz for boys, decreasing to about 125 Hz from the age of 11 to 15 years, and about 270 Hz for girls, descending to about 230 Hz for adult women. Similar values were already published by Peterson and Barney [36] but without distinguishing boys and girls or specifying the age. However, they included average frequencies for the three first formants F1, F2, and F3 of men, women, and children for ten English vowels [36]. Averaging their formant frequencies over all vowels, we find that F1 increases about 14% from a child voice to a woman's voice, and about 33% to a man's voice. The increase is maybe slightly higher for F2, and about 18% and 38% for F3.

Finally, the aged voice presents a new set of characteristics: decreased intensity, breathiness, relatively high pitch (especially for men), lower flexibility, and perhaps trembling [39].

3.1.2. Phonatory settings

The voice can take many different phonatory settings (in addition to those necessary for making the phones of a language). For example, the vocal tract may be shaped to make a dark or bright color or sound nasal. Also interesting are the possibilities of the larynx, which has a great repertoire of voice qualities. Based on numerous studies by several researchers, J. Laver has made a comprehensive discussion and summary of the phonatory settings of the larynx and their relation to the perceived voice quality [40]. He argues for the existence of six basic phonatory settings: *modal voice* and *falsetto* (orthogonal mechanisms), *whisper* and *creak* (combinable with each other and with the first category), and *breathy* and *harsh* voice. Although these are phonatory settings, such vocal qualities may also be provoked by the physical state of the voice such as fatigue, injury, and illness. J. Esling and Al. have later specified the contribution to phonation of the false vocal folds and of the constrictor muscles further above, as summarized in [41]. This helps explaining constricted and unconstricted phonation modes and characterizes harsh, creaky and whispery voices as constricted phonation modes and modal voice, falsetto and breathy voice as unconstricted ones.

3.2. Recording of voice qualities

In addition to considerations of the physiology and the acoustics of the voice, one male and one female actor were recorded saying 10 sentences (in French) while faking different voice qualities depending to their abilities. The voice qualities included soft, tense, breathy, hoarse, whispering, nasal and lisping voices, as well as the voice of an old person, a child, a drunk or the effect of a stuffed nose.

Comparison with their normal voice, which was also recorded (at 48 kHz and 24 bits), gave important spectral information for the transformations, as discussed below.

3.3. Voice Transformations of type and nature

Transformations of type and nature of the voice were grouped into three categories: transformation of physical characteristics of the speaker (size, gender and age), transformation of voice quality (modification of the glottal source to make the voice breathy, whispering, rough, soft, tense, loud etc.), and transformation of speech style (modification of the prosody; liveliness, speech rate, etc.).

3.3.1. Transformation of size, gender and age

While there are general differences between voices of different gender and age, there are also considerable differences within each category. This makes it difficult to determine absolute parameters for successful transformation of the voice, and even though the parameters would be correct, the perception of gender or age may be disturbed by the fact that the speech style does not correspond to the voice. Nevertheless, with the pitch values given in Section 3.1.1 as reference, modification of pitch to change gender and age may simply be achieved by a transposition of the source signal to the given target pitch.

But merely increasing the pitch of a man's voice does only make a man speak in falsetto, or as Mickey Mouse if the spectral envelope is transposing together with f_0 and the harmonics, for instance. The vocal tract should be modified independently by transposing the spectral envelope according to an average of the ratios of the formants of men, women and children given in Section 3.1.1 In order to achieve other voices, such as a teenaged boy or girl, intermediate values were chosen.

In an interactive system, such as TRAX, a transformation plugin presented in Section 3.4, the operator can in addition be given the possibility to optimize the parameters for each voice. In some cases it may indeed be interesting to play with the ambiguity of the gender of the voice and thus choose an intermediate setting, as we did with Céladon's voice when he disguises himself as a woman in the film "Les amours d'Astrée et de Céladon" by E. Rohmer, 2007.

When it comes to aged voices, the characteristics mentioned in Section 3.1.2 are converted into transformations. A convincing old voice can be achieved by the following four measures: trembling is implemented as a slowly fluctuating transposition factor, the pitch is slightly raised (together with the spectral envelope to give a brighter timbre), the speech rate is slowed down, and the f_0 ambitus is decreased. Additionally, breathiness may be added, as described below.

Finally, no literature was found on transformation of size, but we can simply extrapolate our knowledge about gender and age transformation. The approach is intuitive, as when we read a book aloud for a child: raising the pitch and making the vocal tract smaller (transposing the spectral envelope upwards in frequency) make us sound like a small dwarf, and speaking with a low pitch and making the mouth cavity large (downwards spectral-envelope transposition) simulate

the voice of a giant, for instance. Adding breathiness may be an efficient addition to a deep dragon's voice, for instance.

3.3.2. *Whisper*

When we whisper, the vocal folds are separated enough not to vibrate but are still sufficiently close to produce audible turbulence. Recordings showed that the spectral envelope of whisper and voiced speech are comparable at high frequencies (above the estimated VUF) but differ at low frequencies for voiced phones. While the formant frequencies have approximately the same positions, the spectral tilt was flat or even positive below the VUF.

To transform voiced speech to whisper, a source of white noise was therefore filtered by the spectral envelope estimated from the original signal, except for some modification at low frequencies: Firstly, the spectral tilt was neutralized and even inverted below about 3 kHz, depending of the voice. The choice of 3 kHz was an empiric compromise because using the VUF as cut-off frequency for the inversion tended to create audible discontinuities. Secondly, fricatives (unvoiced phones such as /f/, /s/, /θ/, /∫/) should not be touched by this transformation as they depend on turbulence created at constriction further downstream. Since these noisy sounds have the energy concentrated at higher frequencies (in the range 3-6 kHz depending on the sound [24]), preserving the fricatives was indirectly achieved by the measure described above by allowing only to increase the low-frequency spectral tilt.

3.3.3. *Breathy voice*

A breathy phonation is obtained by reducing the force with which the vocal folds are pressed together. The vocal folds vibrate, but the closing movement is not complete or sufficiently soft for air leakage to cause turbulence noise in addition to harmonics. The effect of this on the spectrum is an increasing spectral tilt of the harmonic parts of the signal (i.e., an attenuation of high-frequency harmonics) accompanied by an addition of aspiration noise above about 2 kHz [42].

To render a voice breathy, we proceed in 2 steps: first, the voice must be softened by passing it through a lowpass filter, then noise must be added. Of course, the noise must change with the signal, which is exactly the case for the spectral envelope. An approach similar to that of whisper is thus followed, and the noise is attenuated at low frequencies to avoid it to interfere with the original signal. However, just as with whisper, the fricatives should not be touched. It is therefore important to modulate the lowpass filter with the voicing coefficient. The original, lowpass-filtered signal is then mixed with the whisperlike noise at a ratio that depends on the desired degree of breathiness.

As we have stated at the end of Section 2.2 Rd estimation have been recently implemented into SUPERVP so that it is now also possible to change the vocal quality (e.g. whisper, breathy) by modifying the Rd parameter.

3.3.4. *Transformation of speech style*

Differences between gender and age are also seen in terms of speech style. Speech style is much more difficult to address from a global point of view because it requires a prosodic analysis and processing. It was shown, however, that the dynamic range of the pitch, the pitch ambitus, is a speech-style attribute which varies from speaker to speaker and seems generally greater for children and teenagers than for adults, and even smaller for aged people. Changing the ambitus was efficiently achieved by exaggerating or attenuating the natural variations of f_0 by dynamically transposing the signal in proportion to the log-f_0 deviation from the established median f_0. The median f_0 was chosen rather than the mean f_0 because the median is invariant of the method used for this transformation. Another speech-style attribute is the speech rate. Slowing down the speech to get an aged person, for instance, was done by dilating the signal by some 20 to 50% without changing the pitch or spectral envelope. Changing the ambitus and the speech rate together has surprising effects: dullness may well be achieved from a neutral recording by decreasing the speech rate and the ambitus. Conversely, the opposite transformation of the same neutral recording gives the effect of eagerness.

3.4. ircamTools TRAX

The study presented in the Section 3.3 led to VOICEFORGER, a library dedicated to voice transformations of type and nature based on SUPERVP. In collaboration with the company FLUX which designed the graphical interface, IRCAM, recently proposed the ircamTools commercial plug-in TRAX based on VOICEFORGER. Most of the transformations of TRAX features the interactive design of sound transformations using either real time sound input or sound files loaded into the application. Through the use of an intuitive interface, the effect of all parameter modifications applied to sound transformations can be heard in real time. TRAX is a tool designed for voice but also music

Figure 2. *The ircamTools* TRAX *graphical interface by Flux (See [11] for online demos).*

sonic transformations allowing independent transposition of pitch and timbre (spectral envelope). It offers precise control

over all transformations (transposition, transposition jitter, component remixing, filtering and generalized cross synthesis). It features a creative set of presets. All parameter settings can be saved and recalled.

For single voice sounds there exist presets that allow for high quality transformations of the gender and age of the speaker. These presets can be fine tuned to the specific characteristics of the input voice. For musical sounds an additional mode for transient detection and preservation is available. When transient detection is enabled, the component remixing object allows for the independent remixing of the sinusoid, noise and transient components. Finally, transformations can be stored either as user defined presets or as SUPERVP command lines. Using command lines enables the possibility to apply batch mode transformation to many sound files at once using the settings that have been designed with TRAX.

4. Speaker Conversion

When the desired target voice is specific, it is possible to use voice conversion techniques. The aim of *Speaker Conversion* - a typical application of *voice conversion* technique (VC) - is to modify the speech signal of a source speaker to be perceived as if it had been uttered by a target speaker [43]. It can be of interest on a performative perspective. For instance, it could be used to exchange voice of speakers or singers or to convert an actor's voice towards the voice of an disappeared celebrity or towards the voice of a famous singer. The overall methodology for speaker conversion is to define and learn a mapping function of acoustic features of a source speaker to those of a target speaker. Several approaches have been proposed such as vector quantization [44], neural networks [45] or multivariate linear regression [46] among others statistical methods. One of the most popular statistical method, proposed by Stylianou and al. [47], is based on a *Gaussian Mixture Model* (GMM) that defines a continuous mapping between the features of source and target voices.

Research on speaker conversion have been initiated at IRCAM in [48]. We recently proposed in [49] a new method for spectral conversion called *Dynamic Model Selection* (DMS) based on the use of several models in parallel, assuming that the best model may change over time according to the source acoustic features. We first recall the GMM-based spectral conversion method. Then, we introduce the DMS method in Section 4.2.

4.1. GMM-based spectral conversion

Stylianou and al. [47] proposed to model the source speaker acoustic probability space with a GMM. The cross-covariance of the target speaker with source speaker and the mean of the target speaker were then estimated using least squares optimization of an overdetermined set of linear equations. Kain extended Stylianou's work by modeling directly the joint probability density of the source and target speaker's acoustic space [50]. This joint probability density is estimated on a parallel corpus in which source speaker utterance and target speaker utterance have been temporally aligned. This method allows the system to capture all the existing correlations between the source and target speaker's acoustic features. The conversion is finally performed at each frame n on the basis of the minimum mean-square error (MMSE) : the converted feature vector $\hat{y}_n$ is the weighted sum of the conditional mean vectors $E^y_{k,n}$ in which the weights are the posterior probabilities $p(u_n = k|x_n; \phi_k)$ of the source acoustic feature vector belonging to each one of the mixture components $(u_n = k)$:

$$\hat{y}_n = E[y_n|x_n] = \sum_{k=1}^{K} p(u_n = k|x_n; \phi_k) E^y_{k,n} \qquad (2)$$

The conditional covariance matrix can also be evaluated, giving a kind of confidence measure for the conditional mean vector for each $n \in \mathcal{N}$ which can be use to renormalize the variance of the converted speech parameters. Although this type of method is relatively efficient, conversion performance are still insufficient regarding speech quality: the frame by frame conversion process induces inappropriate spectral parameter trajectories and the converted spectrum can be excessively smoothed. So improvements are still necessary to make it usable for instance for artistic applications which are demanding considering quality.

4.2. Dynamic model selection

In classical speaker conversion methods, a single model is selected during the training step and used for the conversion. This model is selected among others according to the spectral distortion obtained from the conversion of a development corpus or by using methods from the models selection research field. Information Criteria such as BIC [51] have been designed for this purpose. A good model will balance goodness of fit and complexity, so it should have neither a very low bias nor a very low variance. A model with too large a variance due to overparametrization will give poor performance on data different or far from the training data because of the high variance of the local estimators resulting in overfitting. The model undergoes oscillations that are both very large and whose features strongly depend on the exact positions of the points leading to a model with a huge variance and very large response errors. However, the same model will give excellent conversion performances on datas similar or close to the training ones. The *Dynamic Model Selection* (DMS) method proposed in [49] consists of using several models in parallel assuming that the best model may change over time according to the source acoustic features. To do so, a set of potential best models $\mathcal{M}$ including GMMs with increasing number of components is built during the training step. However, the increase of the number of components of a Gaussian mixture is limited by the increasing

complexity of the model due to the large number of parameters associated with the covariance matrices. One way to solve this problem is to use diagonal structures, but the performances are then sacrified because the latter are unable to model the underlying second order statistics.

Mixture of *Probabilistic Principal Component Analyzers* (PPCAs) is a method proposed by Tipping and Bishop [52] to solve the inflexibility of GMMs by performing a pseudo-local *Principal Component Analysis* (PCA) on each mixture component. Modeling covariance structure with a mixture of PPCAs provides an entire range of covariance structures that incrementally includes more covariance information. Mixture of PPCAs can be seen as a more general case of the GMMs for spectral conversion. It can be used in order to define models with increasing number of mixtures while keeping a reasonable model complexity. For the DMS method, several joint models including GMMs with full covariance matrices and mixture of PPCAs are estimated. Then, a set of best potential models - denoted $\mathcal{M}$ - is selected according to the BIC criterion (the best models being the ones with the lowest BIC values).

During the conversion step, at each frame $n \in \mathcal{N}$, the most appropriate model is chosen according to the likelihood of the source datas given each model as

$$\hat{M}_n = \arg\max_{M \in \mathcal{M}} p(x_n|M) \tag{3}$$

with

$$p(x_n|M) = \sum_{k=1}^{K} p(u_n = k)\mathcal{N}(x_n; \bar{\mu}_k^x, \Sigma_k^{xx}) \tag{4}$$

the values of $p(u_n)$ $\bar{\mu}_k^x$, Σ_k^{xx} and K, depending on the model M. In this way, we aim to use a general model with low complexity if the values are far from training data and a more complex and precise model if the source data are closer to training data, leading to a better conversion. An example of model selection along a segment of a speech utterance is given on Figure 3: complex models are used on stable spectrum parts while simpler and general models are used in transition parts.

Subjective tests presented in [49] showed that the method is promising as it can improve the conversion in terms of proximity to the target and quality compared to the method based on a single model. In further work, we will focus on other criteria than the likelihood for the selection of the best model during the conversion. Finally, speaker conversion as been recently implemented into SUPERVP allowing realtime speaker conversion.

5. Transformation of expressivity

We finish this short review of voice transformations at IRCAM by presenting latest reasearch which have been done on the transformation of expressivity in speech. This research has been initiated at IRCAM in [54]. We proposed

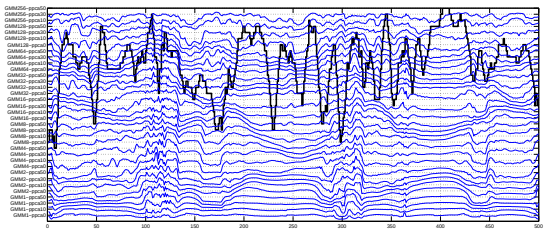

Figure 3. *Example of Dynamic Model Selection along a segment of a speech utterance. On the left the set of potential models. In bold line: selected models at each frame n, in light lines: LSF representation of the source spectral envelope.*

and designed a system of transformation based on Bayesian networks trained on expressive corpuses. This system has yielded interesting results of transformation and its design was accompanied by the development of several descriptors including those concerning the degree of articulation for the expressivity. The study has been continued, which led to a second system that we briefly describe in this review. The general objective is to convert the prosody of a neutral speech into an expressive one. The proposed system, does not require any transcription or phonetic alignment of the input speech since it relies only on the acoustical data. In Section 5.1, we present the expressive corpuses and introduce the prosodic model in 5.2. We then describe the training procedure of the transformation associated to a given expressivity change in the prosodic's parameter space in Section 5.3. The last part details the generation of prosodic trajectories and the implementation of the prosodic transformations. We conclude by giving some perspectives for further work.

5.1. Expressive corpuses

We first describe the expressive corpuses that have been recorded for the training of the models of transformation. Two actors - one man and one woman - have recorded 100 utterances. These utterances were of different lengths with different number of syllables, different size of prosodic phrases and different number of breaks in order to get a broad coverage of various prosodic structures. In this work, only the first 4 basic emotions described by P. Ekman [55] were considered: *joy*, *fear*, *anger* and *sadness*. Thus, different corpuses were recorded with different intensities of emotions. The corpuses were segmented automatically by IRCAMALIGN[56], a French automatic speech segmentation software. This alignment was only used to establish a matching between neutral and expressive prosodic trajectories during the training. On the other hand, an annotation of the prominences was performed on the neutral corpus.

5.2. Prosodic model

Prosodic modeling is achieved on the syllable level. To do so an automatic syllable segmentation has to be performed.

The algorithm is introduced in the next Section 5.2.1. In Section 5.2.2, we present the different prosodic descriptors.

5.2.1. Syllable segmentation

A preliminary step of the prosodic modeling is to segment the speech into syllables. This segmentation is performed automatically from the acoustic data. The syllable detection is based on the Mermelstein algorithm [57]. The detection principle is to identify the energy dips that correspond to the boundaries of syllabic nuclei as illustrated on Fig. 4. However, the detection score was greatly improved by deriving the syllable loudness from the True envelope spectrum [24].

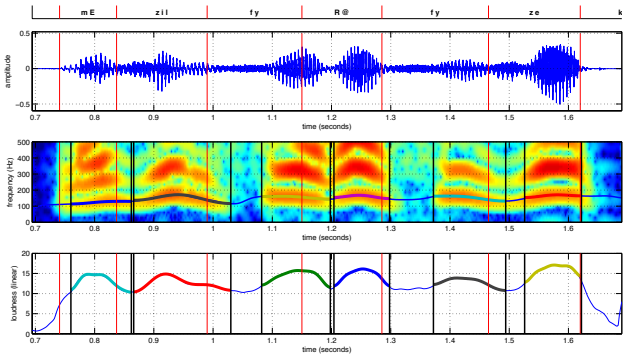

Figure 4. *Detection of syllable nuclei*

5.2.2. Prosodic descriptors

We followed the 5-dimensional model of prosody proposed by Pfitzinger in [58], and extract the following descriptors on a syllable basis:

- *Intonation* characterizes the fundamental frequency curve. It is calculated from the syllable segmentation. Discrete Cosinus Transform (DCT) is used with a high order (7 coefficients) across the syllable. The first three factors are used to model the main speech gestures (mean, slope, curvature). The higher order coefficients are used primarily for modeling the effects of vibrato. We can then analyze the corpuses and characterized them in terms of these descriptors. A Principal Component Analysis (PCA) was made on DCT coefficients on the prominent syllables which are likely to reveal significant prosodic gestures. One can thus infer a dictionary of prototypical forms of the first principal component according to the different emotions as presented in the Figure 5. Registry information provides information on the intra-or extroverted nature of the involved emotion.

- *Speech rate* is based on the measure of the syllable rate and on the ratio between different durations of speech elements, e.g. between the duration of voiced part and unvoiced part, or between duration of active speech compared to breaks.

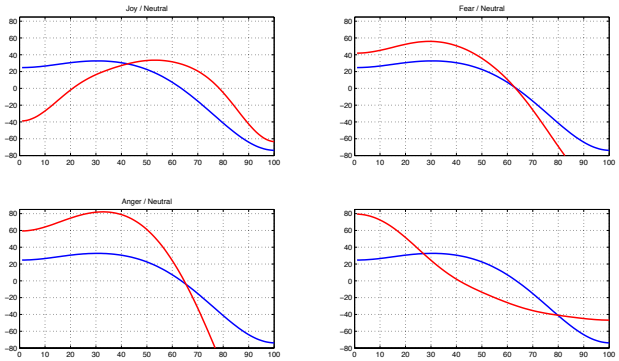

Figure 5. *First principal component on DCT computed on f_0 of prominent syllables according to different emotions. In blue line: neutral speech, in red line: expressive speech*

- *Voice quality* is mainly based on the relaxation index Rd introduced in Section 2.2. Rd measures the relaxed or tensed quality of the glottis to which we added estimates of jitter, shimmer and harmonic to noise ratio (Voiced/Unvoiced Frequency VUF). We also use the algorithm for detecting GCI also presented in Section 2.2. These marks indicate the times of closure of the glottis and allow to compute a measure of jitter. A fast evaluation shows that the neutral and sadness have a small amount of jitter compared to joy and anger.

- *Articulation* is described by estimating the dilatation coefficient between the frequency spectrum of neutral speech compared to the one of expressive speech. This coefficient is estimated by minimizing the log-spectral distance between the original and the dilated spectrums. The value of this coefficient is fixed arbitrarily for the moment and will be automatically estimated in the future.

5.3. Learning transformation functions

The principle of the system is to apply a relative transformation of the prosodic parameters for each syllable. This relative transformation depends on the context of each syllable. Since the corpuses are aligned, it is possible to define transformation vectors between each syllable of the neutral and expressive speech. In the current implementation, two separate transformation vectors are estimated:

- A transformation vector in the joint space of f_0 DCT coefficient and duration

- A transformation vector in the joint space of Rd and jitter parameters.

A context-dependent clustering of these transformation vectors is the performed according to the following contextual features:

- Position of the syllable in the utterance, categorized in beginning, middle, end of utterance
- Position of the syllable in the prosodic group (segment of the speech between 2 breaks)
- Prominent or non-prominent syllable, its kind of prominence, whether it is due to the fact that the syllable is longer than its neighbors or if it is because it has a f_0 peak higher than that of its neighbors.

A *decision tree* learning is used to build vector classes as homogeneous as possible depending on context. The training is done by minimizing the distance between the transformation vectors. Finally, the decision tree will allow to find which transformation vector to apply to a syllable according to its context, even if the context has not been observed in the training corpuses.

5.4. Transformation of expressivity

During the transformation step, vectors to use are selected according to the context of each syllable using the decision tree. To synthesize the trajectories of f_0, the same principle used in HMM-based speech synthesis is used. Dynamic constraints (first derivative and second derivative) are taken into account during the estimation of first DCT coefficient that best explains observation data (maximum likelihood estimation). This can be written as a system of nonlinear equations which can be solved using weighted least squares. This allows to generate a smooth trajectory and finally a sort of gesture on the entire utterance. The same principle is used to generate the trajectory of f_0, Rd and the vowels durations. Jitter and warping are modeled independently as additional effects.

During the transformation of a neutral speech, the speech signal is first segmented into prosodic groups (using voice activity detection) and syllables (using our syllable segmentation algorithm described in 5.2.1). Prominences are then automatically determined using a duration test. Then a f_0 stylization step is done by calculating the DCT coefficients for f_0 on each syllable. The decision tree is then used to determine which transformation to apply according to the context of each syllable. The different curves that are generated are applied using SUPERVP via VOICEFORGER. Among these are transposition curves, dilatation curves for the speech rate, modification curves for Rd which is now integrated into SuperVP and jitter which is simulated by short-term transpositions.

Results depend heavily on the speech to transform. It will be necessary to learn more general transformation models in order to allow several possible strategies of transformation. A subjective evaluation of the system needs also to be done. In the longer term it would be interesting to vary the expressivity along the utterance and therefore to not have discrete categories such as it has been considered until now, which could be achieved by representing the emotions along activation and valence axis.

6. Conclusion

Methods for transformation of gender and age, voice qualities whisper and breathy, and speech style have been presented. With the commercial plug-in TRAX it is now possible to design and apply voice transformations in real-time using an interactive interface. Combined with TTS synthesis, it provides a powerful creation tool which can be use as a performative TTS system. When the voice target is specific, dynamic model selection can be use for speaker conversion. It will soon be implemented into SUPERVP to allow real-time speaker conversion. Finally we presented recent research on the transformation of expressivity which will surely be of interest on a performative perspective in the next years.

References

[1] X. Rodet, "Time-Domain Formant-Wave-Function Synthesis", In *Spoken Language Generation and Understanding*, J.C. Simon, ed., D. Reidel Pub. Co., Dordrecht, Holland. pp. 429-441.

[2] X. Rodet and P. Depalle, "Synthesis by Rule: LPC Diphones and Calculation of Formant Trajectories", In *Proc. ICASSP'85*, Tampa, Fl., march 1985.

[3] X. Rodet, P. Depalle and G. Poirot, "Diphone Sound Synthesis based on Spectral Envelopes and Harmonic/Noise Excitation Functions", In *Proc. 1988 Int. Computer Music Conf.*, Koln, Germany. Sept. 1988. pp. 313-321.

[4] Bennett G. and X. Rodet, Synthesis of the Singing Voice, In *Current Directions in Computer Music Research*, M. V. Mathews and J. R. Pierce, eds., The MIT Press, 1989.

[5] A. Roebel and X. Rodet, "Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation", In *Proc. International Conference on Digital Audio Effects*, Madrid, 2005.

[6] A. Roebel, "Estimation of partial parameters for non stationary sinusoids", In *Proc. International Computer Music Conference (ICMC)*, New Orleans, 2006.

[7] A. Roebel, "A shape-invariant phase vocoder for speech transformation", In *Proc. DAFx-10*, Graz, Austria, 2010.

[8] A. Roebel, "Shape-invariant speech transformation with the phase vocoder", In *Proc. Interspeech 2010*, Makuhari, Japan, 2010.

[9] X. Rodet and A. Lefevre, "The Diphone program: New features, new synthesis methods and experience of musical use", In *Proc. 1997 Int. Computer Music Conf.*, 1997.

[10] G. Peeters and X. Rodet, "SINOLA : A New Method for Analysis/Synthesis using Spectrum Distorsion, Phase and Reassigned Spectrum", In *Proc. International Computer Music Conference*, Pekin, China, Octobre 1999.

[11] http://www.ircamtools.com

[12] G. Beller, C. Veaux, G. Degottex, N. Obin, P. Lanchantin and X. Rodet, "IRCAM Corpus Tools: Système de Gestion de Corpus de Parole", In *Proc.TAL*, 2008.

[13] C. Veaux, G. Beller and X. Rodet, "IrcamCorpusTools: an extensible platform for speech corpora exploitation", In *Proc. LREC*, Marrakech, 2008.

[14] C. Veaux, P. Lanchantin and X. Rodet, "Joint Prosodic and Segmental Unit Selection for Expressive Speech Synthesis",In *Proc. 7th Speech Synthesis Workshop (SSW7)* , Kyoto, Japan, 2010.

[15] P. Lanchantin, G. Degottex and X. Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method", In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010

[16] N. Obin, P. Lanchantin, M. Avanzi, A. Lacheret and X. Rodet, "Toward Improved HMM-based Speech Synthesis Using High-Level Syntactical Feature", In *Proc. Speech Prosody*, Chicago, USA, 2010.

[17] N. Obin, X. Rodet and A. Lacheret, "HMM-based Prosodic Structure Model Using Rich Linguistic Context", In *Proc. Interspeech*, Makuhari, Japan, 2010.

[18] X. Rodet, Y. Potard and J-B. Barrière, "Chant. De la synthèse de la voix chantée à la synthèse en général", In *Rapport de recherche Nž 35*. IRCAM. 1985.

[19] N. Schnell, G. Peeters S. Lemouton, P. Manoury and X. Rodet, "Synthesizing a choir in real-time using Pitch-Synchronous Overlap Add (PSOLA)", In *Proc. ICMC*, Berlin, 2000

[20] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech, In *Speech Communications*, vol. 16, pp. 175-205, 1995

[21] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis", In *Transaction SAP*, 9(1) pp.21-29, 2001.

[22] H. Kawahara, I Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representa- tions using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency- based f0 extraction: Possible role of a repetitive structure in sounds", In *Speech Communication*, 27(3-4):187-207, 1999.

[23] J. Laroche, "Frequency-domain techniques for high-quality voice modification", In *Proc. Int. Conf. on Digital Audio Effects (DAFx)03*, London, UK, 2003.

[24] A. Roebel, "A new approach to transient processing in the phase vocoder", In *Proc. DAFX03*, London, UK, pp344-349, 2003.

[25] N. Bogaards and A. Roebel, "An interface for analysis-driven sound processing", In *Proc. 119th AES Convention*, oct 2005.

[26] G. Degottex, "Glottal source and vocal-tract separation", *Phd thesis UPMC-Ircam*, 2011.

[27] M. Dolson, "The phase vocoder: A tutorial", In *Computer Music Journal*, vol 10, no.4, pp.14-27, 1986.

[28] M. F. Schwartz, "Identification of speaker sex from isolated, voiceless fricatives", In *Journal of the Acoustical Society of America*, vol.43, no. 5, pp. 1178-1179, 1968.

[29] G. Fant, "The LF-model revisited. transformations and frequency domain analysis", *STL-QPSR*, vol. 36, no. 2-3, pp. 119-156, 1995.

[30] G. Degottex, A. Roebel and X. Rodet, "Phase minimization for glottal model estimation", *IEEE Transactions on Acoustics, Speech and Language Processing*, accepted on August 2010.

[31] G. Degottex, A. Roebel and X. Rodet, "Glottal Closure Instant detection from a glottal shape estimate", In *Proc. 13th International Conference on Speech and Computer*, SPECOM, pages 226-231, 2009.

[32] G. Degottex, A. Roebel and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter", In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011

[33] A. de Cheveigné and J. Kawahara, "Yin, a fundamental frequency estimator for Speech and Music", In *Journal of the Acoustical Society of America*, vol. 111, no.4, pp. 1917-1930, 2002

[34] A. Roebel, F. Villavicencio and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order", In *Pattern Recognition Letters*, vol. 28, pp. 1343-1350,2007.

[35] S. Farner, A. Roebel and X. Rodet, "Natural transformation of type and nature of the voice for extending vocal repertoire in high-fidelity applications", In *Proc. AES 35th International Conference*, London, UK, 2009.

[36] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels", In *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175-184, 1952.

[37] K. Wu and D.G. Childers, "Gender recognition from speech. Part I: Coarse analysis", In *Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1828-1840, 1991.

[38] M. Iseli, Y. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source", In *Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2283-2295, 2007.

[39] R. J. Baken, "The aged voice: A new hypothesis", In *Journal of Voice*, vol. 19, no. 3, pp.317-325, 2005.

[40] J. Laver, "The phonetic description of voice quality", *Cambridge studies in linguistics* Cambridge University Press, 1980.

[41] L. D. Bettany, "Range exploration of phonation and pitch in the first six months of life", Master of arts, University of Victoria, 2002.

[42] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", In *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820-857, 1990

[43] M. Abe, S. Nakamura, K. Shikano, and H Kuwabara, "Voice conversion through vector quantization", In *J. Acoust. Soc. Jpn. (E)*, vol. 11, no.2, pp. 71-76, 1990.

[44] M. Abe, S. Nakamura, and H. Kawabara, "Voice conversion through vector quantization", In *Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP'88)*, pp. 655-658, 1988.

[45] M. Narendranath, H. Murthy, S. Rajendran, and B. Yegnanarayan, "Transformation of formants for voice conversion using artificial neural networks", In *Speech Communication*, vol. 16, pp. 207-216, 1995.

[46] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using psola technique", In *Speech Communication*, no 11, pp.175-187, 1992.

[47] Y. Stylianou, O. Cappé , and E. Moulines, "Continuous probabilistic transform for voice conversion", In *IEEE Transactions on Speech and Audio Processing* vol. 6, pp. 131-142, 1998.

[48] F. Villavicencio, A. Roebel and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by True-Envelope estimation", In *Proc. ICASSP*, Toulouse, 2006.

[49] P. Lanchantin and X. Rodet, "Dynamic Model Selection for Spectral Voice Conversion", In *Proc. Interspeech 2010*, Makuhari, Japan, Sept 2010.

[50] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis", In *Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP88)*, pp. 285-288, 1998.

[51] Y.M. Bishop, S.E.Fienberg and P.W. Holland, "Discrete Multivariate Analysis", In *MIT Press Cambridge*, 1975.

[52] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyser", In *Neural Computation*, vol. 11, no. 2, pp. 443-482, 1999.

[53] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory", In *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222-2235, 2007.

[54] G. Beller, "Expresso: Transformation of Expressivity in Speech", In *Proc. Speech Prosody*, Chicago, 2010.

[55] P. Ekman, "The handbook of cognition and emotion", Basic Emotions chapter, John Wiley & Sons, Ltd., 1999.

[56] P. Lanchantin, A. C. Morris X. Rodet and C. Veaux, "Automatic Phoneme Segmentation With Relaxed Textual Constraints", In *Proc. LREC'08* Proceedings, Marrakech, Marocco, 2008.

[57] P. Mermelstein, "Automatic segmentation of speech into syllabic units", In *Journal of the Acoustical Society of America*, vol. 58, pp. 880-883, 1975.

[58] H. R. Pfitzinger, "Five dimensions of prosody: intensity, intonation, timing, voice quality, and degree of reduction", In *Proc. Speech Prosody*, paper KN2, 2006.