Voice quality transformation using an extended source-filter speech model

Stefan Huber, Axel Roebel
Sound Analysis/Synthesis Team, IRCAM-CNRS-UPMC STMS, 75004 Paris, France
axel (dot) roebel (at) ircam (dot) fr

ABSTRACT

In this paper we present a flexible framework for parametric speech analysis and synthesis with high quality. It constitutes an extended source-filter model. The novelty of the proposed speech processing system lies in its extended means to use a Deterministic plus Stochastic Model (DSM) for the estimation of the unvoiced stochastic component from a speech recording. Further contributions are the efficient and robust means to extract the Vocal Tract Filter (VTF) and the modelling of energy variations. The system is evaluated in the context of two voice quality transformations on natural human speech. The voice quality of a speech phrase is altered by means of re-synthesizing the deterministic component with different pulse shapes of the glottal excitation source. A Gaussian Mixture Model (GMM) is used in one test to predict energies for the re-synthesis of the deterministic and the stochastic component. The subjective listening tests suggests that the speech processing system is able to successfully synthesize and arise to a listener the perceptual sensation of different voice quality characteristics. Additionally, improvements of the speech synthesis quality compared to a baseline method are demonstrated.

1. INTRODUCTION

In this paper we present a method to transform the deterministic and stochastic part of the glottal excitation source. The main motivation of the following paper is the presentation of an improved method for coherent modifications of the glottal pulse shape. The glottal pulse shape is generally accepted to reflect different phonation types of human voice production [1] and different voice qualities being strongly related to the vocal effort [2]. The terminology used in the following is describing the lax-tense dimension of voice quality [3] distinguishing tense (pressed), modal (normal), and relaxed (breathy) voice qualities [4].

Recent research in the speech community has notably improved the speech synthesis quality by explicitly modelling the deterministic and stochastic component of the glottal excitation source [5, 6]. Advanced source-filter decomposition strategies as in [7–9] address finer details defined by extended voice production models for human speech. These approaches analyze an extended feature set to model their transformation and synthesis. The extended feature set consists of: the VTF, the glottal pulse positions and shapes, the energies, and a random component described by spectral and temporal envelopes.

In this paper we present a novel speech analysis and synthesis system extending the source-filter model of [9]. The extension is based on using a DSM and further processing means. The deterministic part is estimated and subtracted from a speech signal to extract the stochastic part [10]. The proposed system separately models the stochastic and deterministic components. It does therefore not correspond to the classical source-filter model. The contribution of the following research and the advancements compared to the baseline method lies in the extended means to estimate the unvoiced stochastic component, to robustly extract the VTF and to handle the variations in energy and signal behaviour implied with glottal source transformations.

The paper is organized as follows. Section 2 presents the novel speech framework. Section 3 discusses the aspects of voice quality transformation. Section 4 introduces the baseline state-of-the-art speech system. Section 5 presents a subjective evaluation based on a listening test of natural human speech. Section 6 concludes with the findings studied in this paper.

2. THE EXTENDED SOURCE-FILTER MODEL

The proposed speech analysis and synthesis system is designed for the utilization in advanced voice transformation and voice conversion applications. It is denoted **PSY** for **P**arametric **S**peech analysis, transformation and **SY**nthesis.

2.1 Voice production model

PSY operates upon the following generic interpretation of the human voice production in the time domain:

$$s(n) = u(n) + v(n) = u(n) + \sum_i g(n, P_i) * \delta(n - P_i) * c(n, P_i) \quad (1)$$

The speech signal $s(n)$ is represented by means of a stochastic (unvoiced) component $u(n)$ and a deterministic (voiced) component $v(n)$. The deterministic component contains the sequence of glottal pulses located at the time positions P_i, each representing a Glottal Closure Instant (GCI) with index i. Each glottal pulse is represented by the glottal flow derivative $g(n, P_i)$. The latter is convolved with a Dirac impulse at the GCI P_i and the VTF that is active for the related position $c(n, P_i)$. The Liljencrants-Fant (LF) model [13] is used to synthesize each $g(n, P_i)$. The LF model is parameterized by a scalar shape parameter R_d [14, 15]. Changing R_d continuously from lower to

higher values will allow changing the LF pulse shape on a continuum from tense to relaxed voice qualities.

For being able to make spectral domain manipulations the speech signal model given in equ. 1 is processed in the spectral domain using the Short-Time Fourier transform (STFT). For brevity the coverage of a few consecutive glottal pulses $g(n, P_j)$ will be denoted as $g_s(n) = \sum_j g(n, P_j) * \delta(n - P_j)$ in the following. The summation over the GCI index j is set to comprise a signal segment of a few glottal pulses being covered by the Hanning window $w_h(n)$ of the STFT. Each pulse position is related to a slightly different VTF being supposed to be minimum phase [12]. The glottal pulse shape and the VTF are assumed to not change within the window and are given approximately by the corresponding parameters in the window center.

We further assume that the filtering processes implied by each convolutional operation between the signal components of equ. 1 involves impulse responses that are shorter than the window length. The STFT of the speech signal is then given by

$$S(\omega, m) = U(\omega, m) + V(\omega, m) \quad (2)$$
$$= U(\omega, m) + G(\omega, m) \cdot H(\omega, m) \cdot C(\omega, m) \quad (3)$$

The STFT frame m is the position of the window center and ω is the frequency variable of the Discrete-Time Fourier Transform (DTFT). For brevity the dependency of all signal spectra with respect to m will be dropped in the following. $U(\omega)$ and $V(\omega)$ are the DTFT of the windowed voiced and unvoiced signals from equ. 1 assuming that g and c and the corresponding DTFT spectra $G(\omega)$ and $C(\omega)$ are quasi-stationary within the window. $H(\omega)$ is the spectral representation of the windowed Dirac impulse sequence $\delta(n - P_i)$. The radiation filter at lips and nostrils level $R(\omega)$ [11] is not explicitly present in the PSY model since it is implicitly contained in the glottal flow derivative $g(n)$ and the unvoiced component $u(n)$.

2.2 Glottal source synthesis and VTF extraction

The LF shape parameter R_d is estimated by the means proposed in [16, 17]. Each GCI is estimated by the method described in [18] and assigned the closest R_d value which is estimated on the STFT time grid. The spectral envelope sequence $\mathcal{T}_{sig}$ is estimated on the input signal $s(n)$ using the True Envelope estimator of [19]. Another spectral envelope sequence $\mathcal{T}_g$ is estimated on the synthesized glottal flow derivative sequence $g_s(n)$. The extraction of the VTF $C(\omega)$ is obtained by dividing $\mathcal{T}_{sig}$ by $\mathcal{T}_g$. The utilization of $\mathcal{T}_g$ in the full-band division is required to suppress the spectral ripples occurring for higher R_d values [15, 20].

2.3 Estimation of the unvoiced stochastic part

The separation of a speech signal $s(n)$ into the contributions of the voiced $v(n)$ and the unvoiced $u(n)$ part is based on the calculation of a residual of a sinusoidal model [21]. The following algorithmic step estimate a) the unvoiced residual $u_{res}(n)$ by deleting sinusoidal content from $s(n)$, b) $u_{HP}(n)$ by high-pass filtering $u_{res}(n)$, c) the unvoiced signal $u(n)$ by scaling $u_{HP}(n)$ in energy.

a) Re-Mixing with De-Modulation: This approach aims to simplify the sinusoidal detection by de-modulating the fundamental frequency F_0 contour and the Hilbert amplitude envelope $\mathcal{H}$ from $s(n)$. The original F_0 contour of $s(n)$ is warped to become flat by means of time varying re-sampling using as target F_0' the mean of the original F_0. The re-sampling operation changes locally and globally the time duration of all signal features. The effect will be inverted after the extraction of the residual. The varying amplitude contour of $s(n)$ is demodulated by means of dividing the signal by its smoothed Hilbert transform $\mathcal{H}(s(n))$ similar as in [5,23]. The smoothing kernel is simply the Hanning window of duration $4/F_T$. This optimally removes all envelope fluctuations that are related to the deterministic components. The resulting signal $s_{flat}(n)$ is flat in amplitude envelope and F_0 facilitating the detection of sinusoids following [21]. It avoids even for relatively high harmonic numbers the energy shift between voiced and unvoiced components [22]. The sinusoidal content is subtracted from $s_{flat}(n)$ and the demodulation steps are inverted so that the original AM-FM modulation is recreated. This generates the unvoiced residual signal $u_{res}(n)$.

b) Below F_{VU} filter: Informal tests confirm that not all sinusoidal content could be precisely estimated and deleted in the frequency band below the Voiced / Unvoiced Frequency boundary F_{VU} [24]. The F_{VU} estimation is based on the signal interpretation splitting the spectrum into two bands. The lower frequency band below the F_{VU} is determined by the voiced component $V(\omega)$. The unvoiced component $U(\omega)$ is located above the F_{VU}. A high pass filter is applied to delete remaining sinusoidal content from $U_{res}(\omega)$ below F_{VU}. The filters cut-off frequency f_c equals the estimated F_{VU} per STFT frame m. A gain of 1 is set in the filters passband equalling the stochastic frequency band $\omega > \omega_{VU}$. A linear ramp with a slope of m_{HP}=-3 dB per octave defines the high pass filtering in the filters stopband. The latter equals the deterministic frequency band $\omega < \omega_{VU}$. The experimental findings show that a heuristically defined threshold of m_{HP}=-3 dB approximates reasonably close enough the desired sinusoidal cancellation in the high pass filtered unvoiced signal $u_{HP}(n)$.

c) Scale to $\mathcal{T}_{sig}$ level: The sinusoidal detection of step a) may be erroneous for some signal segments such as fast transients. The heuristic adaptation of step b) cannot be exact for all cases. The scaling described in equ. 4 minimizes the difference between the envelope $\mathcal{T}_{unv}$ of the stochastic component $U_{HP}(\omega)$ and the envelope $\mathcal{T}_{sig}$ of the signal spectrum $S(\omega)$ above F_{VU} up to the Nyquist frequency F_{nyq}. The DFT bins found closest to the frequencies F_{nyq} and F_{VU} are denoted as k_{nyq} and respectively k_{VU}.

$$\eta = \frac{1}{k_{nyq} - k_{VU}} \sum_{k=k_{VU}}^{K=k_{nyq}} \left(\mathcal{T}_{sig}^{dB}(k) - \mathcal{T}_{unv}^{dB}(k) \right)$$
$$\mathcal{T}_{unv}^{w} = \mathcal{T}_{unv} \cdot (1 - k_{VU}/k_{nyq}) \cdot 10^{\eta/20} \quad (4)$$

η equals the mean difference in dB between $\mathcal{T}_{sig}$ and the spectral envelope $\mathcal{T}_{unv}$.The scaling of $\mathcal{T}_{unv}$ is weighted by the time-varying ratio of F_{VU} versus F_{nyq} as a regularization term to avoid a too high energy scaling. The multiplication of a white noise spectrum with $\mathcal{T}_{unv}^{w}(\omega)$ synthesizes with the STFT the unvoiced signal $u(n)$.

2.4 GMM-based F_{VU} prediction

The spectral fading synthesis presented in the following section 2.6.2 requires a transformed F'_{VU} value, with the operator $'$ indicating a transformation. F'_{VU} is predicted using a modified GMM approach detailed in [17, 25, 26]. The GMM model $\mathcal{M}$ is trained on the voice descriptor set d=[R_d,F_0,$H1$-$H2$,E_{voi},E_{unv}] and the reference value $r = F_{VU}$. $H1$-$H2$ refers to the amplitude difference in dB of the first two harmonic sinusoidal partials. E_{voi} and E_{unv} are the Root-Mean-Square (RMS) based energy measures of the voiced and unvoiced signal parts which will be introduced in the following section. The prediction function

$$F(d) = \sum_{q=1}^{Q} p_q^d d \cdot [\mu_q^r + \Sigma_q^{rd} \Sigma_q^{dd^{-1}} (d - \mu_q^d)] \tag{5}$$

is derived from $\mathcal{M}$ by the definition of equ. 5, with Q=15 being the number of utilized Gaussian mixture components. An initial F^p_{VU} value is predicted from $F(d)$. An error GMM model $\mathcal{M}_{err}$ is trained on the modelling error

$$\epsilon_M = \sqrt[2]{(F_{VU} - F^p_{VU})^2} \tag{6}$$

serving as reference value r_ϵ=ϵ_M, and on the voice descriptor set d. The transformed descriptor counterpart d' contains the original F_0 contour but transformed values for the remaining voice descriptors: d'=[R'_d,F_0,$H'1$-$H'2$,E'_{voi}, E'_{unv}]. The GMM-based modelling to predict a F'_{VU} contour from the feature sets d and d' is described by:

$$F^p_{VU_\mu} = \mathcal{M}\left(F(d)\right) \tag{7}$$

$$F^{'p}_{VU_\mu} = \mathcal{M}\left(F(d')\right) \tag{8}$$

$$F^p_{VU_\sigma} = \mathcal{M}_{err}\left(F_{err}(d)\right) \tag{9}$$

$$F^{'p}_{VU_\sigma} = \mathcal{M}_{err}\left(F_{err}(d')\right) \tag{10}$$

$$F'_{VU} = F^{'p}_{VU_\mu} + (F_{VU} - F^p_{VU_\mu}) \cdot F^{'p}_{VU_\sigma} / F^p_{VU_\sigma}. \tag{11}$$

Each trained model pair $\mathcal{M}$ and $\mathcal{M}_{err}$ is utilized to predict via their derived prediction functions F and F_{err} the mean prediction value $F^p_{VU_\mu}$ ($F^{'p}_{VU_\mu}$) and the predicted standard deviation $F^p_{VU_\sigma}$ ($F^{'p}_{VU_\sigma}$) from descriptor set d (from the transformed set d'). The true prediction value would equal $F^p_{VU_\mu}$ if no model error occurs: ϵ_M=0. The calculation of F'_{VU} from the transformed d' and the original voice descriptor set d is defined by equ. 11. It evaluates the difference between the original F_{VU} and the predicted $F^p_{VU_\mu}$ value. The difference result is normalized by the ratio of the original and transformed standard deviations $F^{'p}_{VU_\sigma}$ and $F^p_{VU_\sigma}$ of the modelled data distribution, and corrected by the transformed predicted mean value $F^{'p}_{VU_\mu}$.

2.5 Energy modelling

2.5.1 Energy maintenance

A simple RMS measure F_{RMS} evaluates the effective energy value E on the linear amplitude spectrum A_{lin}=$|Y(\omega)|$ of any arbitrary signal spectrum $Y(\omega)$. The RMS energy measures are estimated in PSY as defined in equ. 12:

$$\begin{aligned} F_{RMS}(A_{lin}, k) &= \sqrt{1/K \cdot \Sigma_k \left(A_{lin}(k)^2\right)} \\ E_{sig} &= F_{RMS}(|S(\omega)|) \\ E_{unv} &= F_{RMS}(|U(\omega)|) \\ E_{voi} &= E_{sig} - E_{unv} \end{aligned} \tag{12}$$

E_{sig} and E_{unv} reflect the RMS energies measured on the signal $S(\omega)$ and the unvoiced component $U(\omega)$. E_{voi} is expressed as their difference to represent the RMS energy of the voiced component $V(\omega)$. A transformed $R^{'gci}_d$ contour causes an altered energy value E'_{voi} measured on the transformed voiced part $V'(\omega)$. The high (low) pass filtering applied to $U(\omega)$ ($V(\omega)$) explained in section 2.6.2 generates as well an energy change. The energy re-scaling to the original energy measures defined by equ. 13 ensures that their energy is maintained:

$$\begin{aligned} E_{voi} &= F_{RMS}(|V(\omega)|) \qquad E_{unv} = F_{RMS}(|U(\omega)|) \\ E'_{voi} &= F_{RMS}(|V'(\omega)|) \qquad E'_{unv} = F_{RMS}(|U'(\omega)|) \\ V'(\omega)\cdot &= E_{voi}/E'_{voi} \qquad U'(\omega)\cdot = E_{unv}/E'_{unv} \end{aligned} \tag{13}$$

2.5.2 GMM-based energy prediction

The original voice descriptor set D_E consists of the voice descriptors D_E=[$R_d, F_0, F_{VU}, H1$-$H2$]. The transformed voice descriptor $H'1$-$H'2$ is measured on the magnitude spectrum of $|S'(\omega)|$ in dB. The predicted F'_{VU} value is retrieved from the signal $S'(\omega)$ and the GMM model of section 2.4. The original and not transformed voice descriptor F_0 is added to the energy modelling due to its high correlation with the other voice descriptors. The manually transformed R'_d, the re-estimated $H'1$-$H'2$, the predicted F'_{VU} and the original F_0 descriptors define the transformed voice descriptor set $D'_E = [R'_d, F_0, F'_{VU}, H'1$-$H'2]$. Each energy model receives for training its corresponding reference feature R defined in equ. 11. The energy models $\mathcal{M}^{voi}$ and $\mathcal{M}^{unv}$ are used via their functions F^{voi} and F^{unv}, along with their corresponding error models $\mathcal{M}^{voi}_{err}$ and $\mathcal{M}^{unv}_{err}$ and error functions F^{voi}_{err} and F^{unv}_{err} to predict the RMS-based energy measures E^p_{voi} and E^p_{unv}.

2.6 Synthesis

2.6.1 Time domain mixing

The straight-forward mixing in the time domain adds the synthesized unvoiced stochastic waveform $u(n)$ to the synthesized voiced deterministic waveform $v(n)$. The time domain mixing operates thus full-band without any restriction on the signal bandwidth. It will be evaluated in section 5.1 together with the GMM-based prediction and scaling of the voiced and unvoiced signal energies.

2.6.2 Spectral fading synthesis

The PSY synthesis variant "Spectral fading" is designed to handle voice quality transformations by suppressing possibly occurring artefacts. A short summary discusses here the impact of R_d on the spectral slope required to understand the motivation for the spectral fading synthesis presented in this section. The spectral slope is strongly correlated with R_d. Altering R_d affects the spectral slope. References to an extensive analysis of the spectral correlates of R_d can be found in [14, 15, 20, 27, 28]. A more relaxed voice quality is reflected by higher R_d values and is related to a sinusoidal-like glottal flow derivative which generates higher spectral slopes. A more tense voice quality is parameterized by lower R_d values and relates to an impulse-like glottal flow derivative which produces lower spectral slopes. A lower (higher) spectral slope indicates that more

(less) sinusoidal content can be observed in higher frequency regions. The voice quality transformation to change an original speech recording having a modal voice quality to a more tense voice character has to extend the quasi-harmonic sequence of sinusoidals above the F_{VU}. Contrariwise, a transformation to a more relaxed voice quality needs to reduce the sinusoidal content. A modification of the glottal excitation source required for voice quality transformations implies a F_{VU} modification. The altered F'_{VU} frequency has to be naturally represented by properly joining the voiced $V(\omega)$ and unvoiced $U(\omega)$ signal components. The transformation of the original R_d^{gci} contour used to extract $C(\omega)$ introduces an energy variation in the re-synthesis of a transformed $V'(\omega)$. However, even with the energy maintenance of section 2.5 the alteration of a modal to a very tense voice quality may result into sinusoidal content being of higher energy than the noise part at F_{nyq}. This sets $F'_{VU} = F_{nyq}$ and causes audible artefacts. F'_{VU} is therefore predicted using the method described in section 2.4. Additionally, the spectral fading method employs two spectral filters to cross fade $V(\omega)$ and $U(\omega)$ around F'_{VU}. The spectral band around F_{VU} is comprised of a mix of both deterministic $V(\omega)$ and stochastic $U(\omega)$ signal content. A low pass filter P_L fades out the voiced part $V(\omega)$ and a high pass filter P_H fades in the unvoiced part $U(\omega)$ with increasing frequency. The linear ramps with a slope of m_{LP}=-96 dB and m_{HP}=-48 dB per octave define the steepness of both filters. A higher value is chosen for m_{LP} since the F'_{VU} prediction may be very high for very tense voice qualities. A less steep fade out filter would not be effective enough.

3. VOICE QUALITY TRANSFORMATION

The study of [29] on the Just Noticable Differences (JND) of human auditory perception reports that changes in higher (lower) value regions of the Open Quotient OQ (the asymmetry coefficient α_m) require longer distances of ΔOQ ($\Delta\alpha_m$) to arise the sensation of a voice quality change in the perception of a listener. We spread according to that experimental results the original R_d^{gci} contour into several $R_d^{'gci}$ contours with positive and negative offsets covering the complete R_d range such that lower ΔR_d steps are placed in lower and higher ΔR_d steps in higher R_d value regions. One example is illustrated in fig. 1 on the phrase employed for the evaluation in section 5. Table 1 shows the mean R_d^{μ} values of the original R_d contour with index 0, and respectively 3 positive and 3 negative μ values for each voice quality change. $R_d^{\sigma^2}$ lists their variance σ^2. It increases with increasing R_d to reflect the objective of having to apply higher ΔR_d steps with higher R_d values. The R_d mean difference column ΔR_d^{μ} reflects the mean ΔR_d steps measured between each row index on the R_d^{μ} values to show that also the mean difference increases with increasing R_d^{μ} from a tense to a relaxed voice quality.

4. BASELINE METHOD SVLN

The method called "**S**eparation of the **V**ocal tract with the **L**iljencrants-Fant model plus **N**oise" detailed in [9, 30, 31]

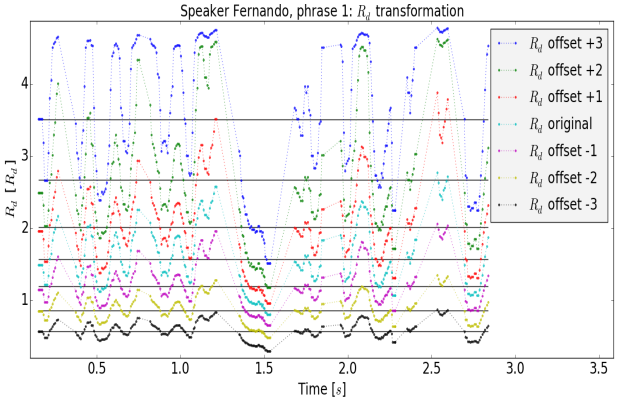

Figure 1: *Generated $R_d^{'gci}$ contour examples*

Voice quality (index)	R_d^{μ}	$R_d^{\sigma^2}$	ΔR_d^{μ}
Very relaxed (+3)	3.5109	0.9031	-0.8397
Relaxed (+2)	2.6711	0.7825	-0.6597
Modal to relaxed (+1)	2.0114	0.3631	-0.4442
Modal (original) (0)	1.5673	0.1937	
Tense to modal (-1)	1.1936	0.0941	-0.3737
Tense (-2)	0.8601	0.0341	-0.3335
Very tense (-3)	0.5704	0.0154	-0.2898

Table 1: *R_d mean, variance and mean difference values*

represents the baseline method on whose means the proposed system PSY is build upon. The main differences are the VTF representation, the energy model and the estimation of the stochastic component. SVLN constructs the latter by high pass filtering white noise, applying an amplitude modulation parameterized by the glottal pulse sequence, and cross fading between consecutive synthesized noise segments. The gain σ_g measures the energy level at F_{VU} while analysis to control the stochastic energy at the synthesis step. SVLN synthesizes glottal pulses with the LF model in the spectral domain to extract $C(\omega)$ below F_{VU}. The VTF above F_{VU} is taken from the signals spectral envelope. SVLN facilitates advanced pitch transposition or voice quality transformations while maintaining a high synthesis quality [9, 32].

5. EVALUATION

The evaluation section presents the results of two listening tests conducted on natural human speech of the Hispanic speaker "Fernando" speaking French. The voice quality assessment examines how well both synthesis systems are able to produce different voice quality characteristics. Test participants were asked to rate different synthesized voice qualities according to the same indices as in table 1. Each phrase is rated as well on their synthesis quality according to the Mean Opinion Scale (MOS).

The baseline method SVLN of section 4 and the proposed method PSY of section 2 received the same features R_d^{gci}, F_0 and F_{VU} as pre-estimated input to analyze their corresponding VTF $C(\omega)$. Please note that SVLN requires to smooth the voice descriptor contours. Due to the energy measure at F_{VU} it cannot handle value changes varying too

quickly in short-time segments [30]. For this test a median smoothing filter covering 100 ms was applied.

5.1 Manual R_d offsets and time domain mixing

A preliminary listening test has been conducted by 6 sound processing experts internally in the laboratory. The listening test is available online via: Manual offset test [1] .

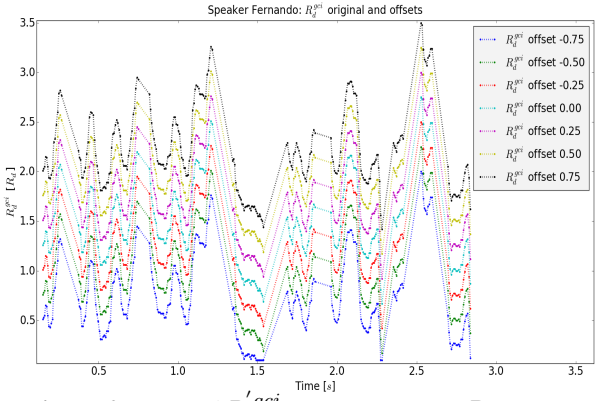

Figure 2: *Manual $R_d^{'gci}$ offsets, step size R_d±0.25*

Fig. 2 depicts the original R_d^{gci} contour in the middle shown in cyan colour, and six additional $R_d^{'gci}$ contours. Each positive and negative mean offset constitutes an empirically determined R_d offset of R_d±0.25 to the previous contour in its respective direction. The offset amount was chosen such that an R_d^{gci} offset contour reaches an R_d range border [0.1 5.0]. In this example the R_d^{gci} offset -0.75 saturates around ∼1.50 seconds on the lower R_d border.

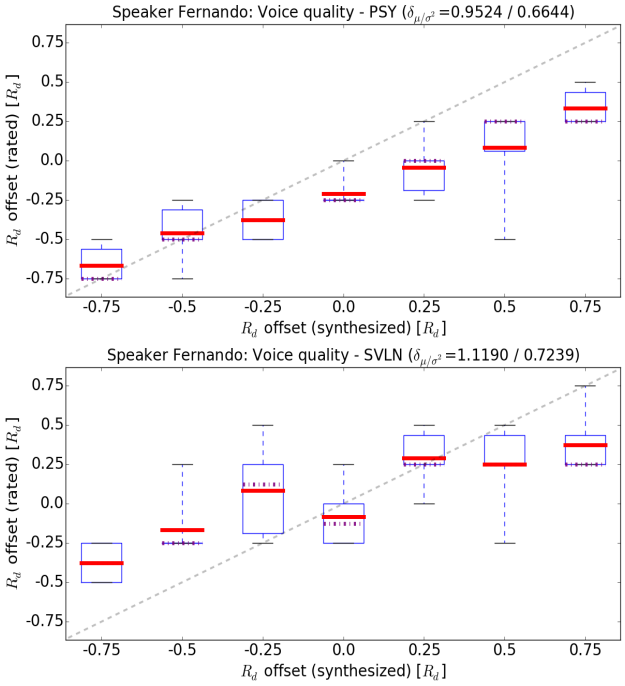

Figure 3: *Voice quality ratings - TD mixing*

Fig. 3 depicts the voice quality ratings for both speech systems. The horizontal grey lines at both ends (whiskers) are set to show the minimum and maximum value for each evaluation. The horizontal red (violet) lines reflect the mean (median) voice quality ratings of all participants per test phrase. The dialog grey dashed line exemplifies their ideal placement if each test participant would have been able to perceptually associate each synthesized voice quality example to its corresponding voice quality characteristic. The mean deviation value δ_μ=0.95 for PSY expresses the disagreement of the listeners, being ideally δ_μ=0.00. PSY received very low mean deviation δ_μ values for more tense voice qualities. The stronger the original modal voice quality is transformed towards a more relaxed voice quality the less well could the participants identify its perceptual sensation. Drawing a regression line through each mean value shown in red horizontal lines per rated R_d offset would result in a less step line than the ideal one depicted as grey dashed line. A higher mean deviation value δ_μ=1.12 as compared to PSY is shown for the baseline method SVLN in fig. 3. It indicates that the listeners could less well capture the different synthesized voice qualities and associate them with the corresponding offset indices. Clear voice quality associations can be concluded for both systems.

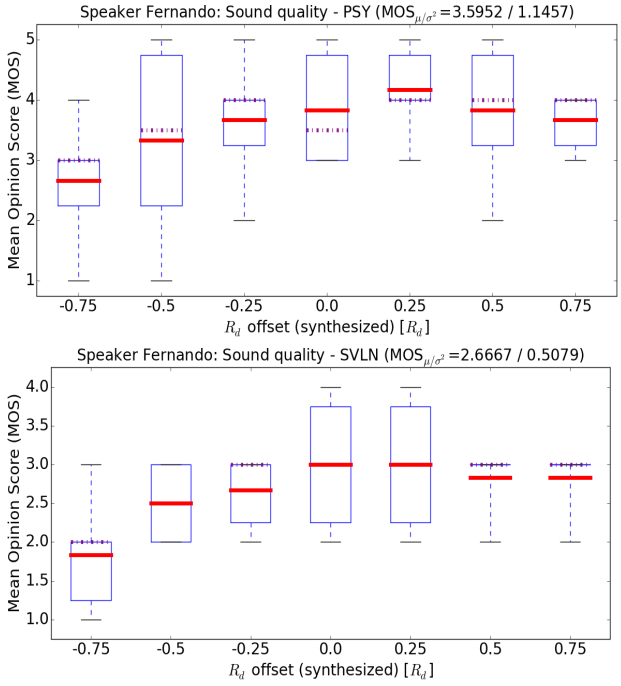

Figure 4: *MOS synthesis quality ratings - TD mixing*

The MOS synthesis quality result are shown in fig. 4. PSY exhibits partially highest ratings up to an excellent synthesis quality of 5 for all but the very tense and very relaxed voice quality characteristics with the R_d offsets ±0.75. Contrariwise, the voice qualities very tense and tense are partially rated with the lowest MOS synthesis quality poor. The mean synthesis quality MOS$_\mu$=2.67 of SVLN is comparably lower than MOS$_\mu$=3.60 for PSY. The very tense voice quality of SVLN received comparably lower MOS ratings than its other synthesized R_d offsets. Stronger voice quality changes are assessed with less good MOS synthesis qualities for both systems. PSY received in general a lower deviation from the true voice quality rating and a higher MOS synthesis quality compared to SVLN.
Fig. 5 illustrates the voice quality and the MOS synthesis quality ratings for the PSY synthesis variant using an

[1] Speaker Fernando: http://stefan.huber.rocks/phd/tests/RdMisterF/

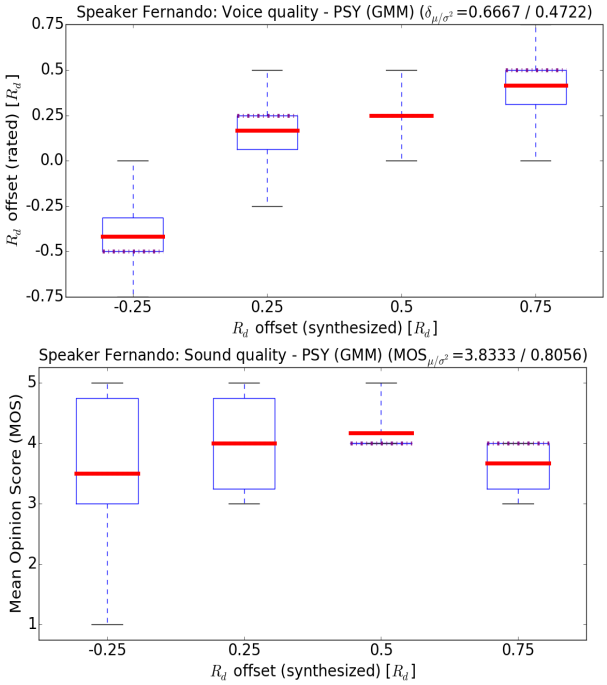

Figure 5: *Test results - PSY energy scaling*

additional energy scaling. The voiced $V(\omega)$ and unvoiced $U(\omega)$ component are scaled by the respective RMS energies predicted from a dedicated GMM energy model for each part. Please note that the two R_d offsets -0.75 for a very tense and -0.50 for a tense voice quality had to be excluded from the test for PSY (GMM). The predicted RMS energy contours resulted into amplitudes in the time domain being outside the valid range [-1 1]. In general it can be observed that the GMM-based energy scaling of PSY received roughly similar voice and MOS synthesis quality ratings as the standard PSY method. This suggests that the GMM predicted energy contours for the voiced $V(\omega)$ and unvoiced $U(\omega)$ parts do neither increase nor decrease the synthesis quality and the voice quality characteristic to a significant extent.

Method	Δ VQ$_\mu$	Δ VQ$_{\sigma^2}$	MOS$_\mu$	MOS$_{\sigma^2}$
PSY	0.9524	0.6644	3.5952	1.1457
PSY (GMM)	0.6667	0.4722	3.8333	0.8056
SVLN	1.1190	0.7239	2.6667	0.5079

Table 2: *Voice quality (VQ) and MOS sound quality*

Table 2 summarizes the mean deviation Δ VQ$_\mu$ and its variance Δ VQ$_{\sigma^2}$ from the optimal voice quality rating in the first two columns. The corresponding mean and variance of the MOS sound quality ratings are listed in the last two columns. The three synthesis approaches PSY time domain mixing in the first row, PSY time domain mixing using the additional GMM energy scaling of section 2.5.2, and the baseline method SVLN are compared. The lower VQ and higher MOS values for PSY (GMM) are partially a result of having omitted the two voice quality transformations towards a tense and very tense voice quality. The expectation for these two omitted test cases is that they would have decreased the good test results for PSY (GMM).

5.2 Transformed $R_d^{'gci}$ contours and spectral fading

The PSY spectral fading synthesis variant presented in 2.6.2 requires the F_{VU} prediction of section 2.4. An example is depicted in fig. 6. The transformed $R_d^{'gci}$ contours and the

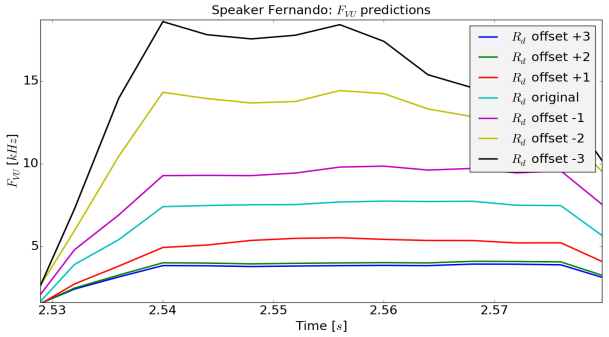

Figure 6: *F_{VU} prediction excerpt for PSY synthesis*

original R_d^{gci} contour of fig. 1 were employed by both systems for synthesis. Following the voice production model of equ. 2.1, a transformed glottal pulse $G'(\omega)$ leads to a transformed reconstructed signal $S'(\omega)$. The unvoiced component $U(\omega)$ remains unmodified. 11 participants rated each speech phrase by SVLN and PSY. Please note that the PSY energy prediction variant is due to the too huge scaling for tense voice qualities omitted. The listening test is available online via: Transformed $R_d^{'gci}$ test [2].

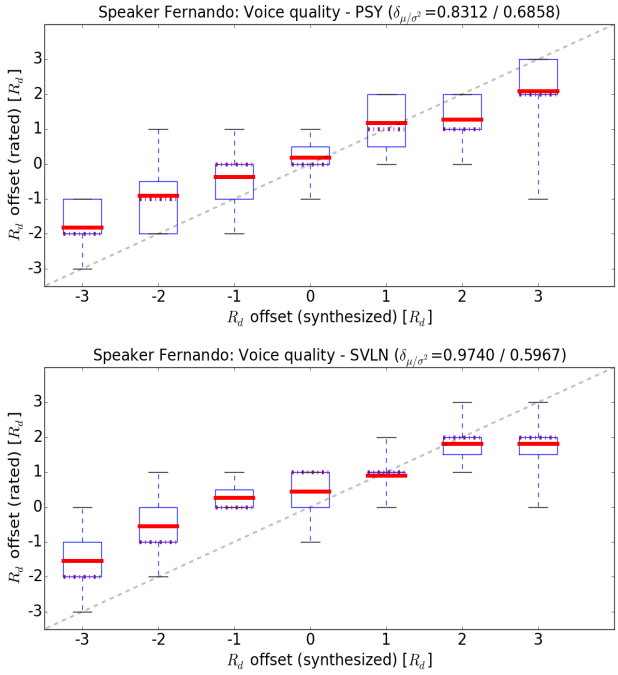

Figure 7: *Voice quality ratings - Spectral fading*

Fig. 7 shows again the voice quality ratings for both speech systems. The mean deviation value δ_μ=0.83 for PSY is lower than the corresponding δ_μ=0.97 for the SVLN. Clear voice quality associations can be concluded for both systems following closely the ideal dashed line. The deviations increase with higher transformations.

[2] Speaker Fernando: http://stefan.huber.rocks/phd/tests/vqMisterF/

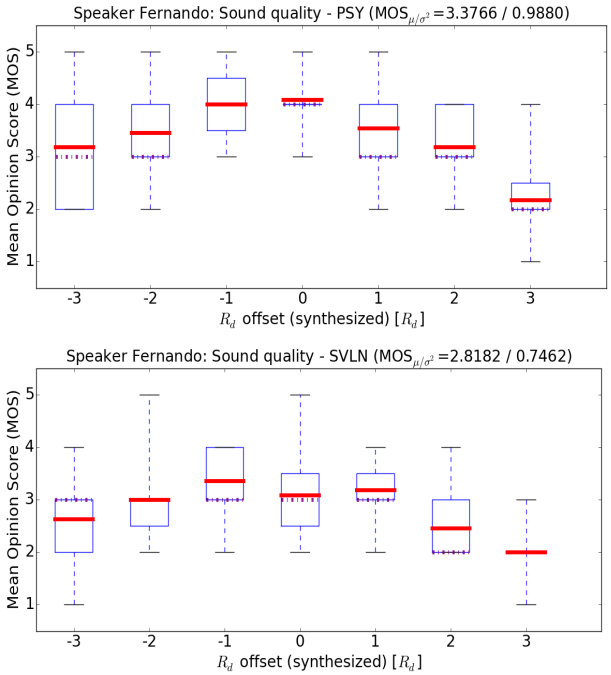

Figure 8: *MOS synthesis quality ratings - Spectral fading*

The MOS synthesis quality evaluation for PSY shown in fig. 8 exhibits partially highest ratings up to an excellent synthesis quality of 5 for all but the "relaxed" and "very relaxed" voice quality characteristics with index +2 and +3. The evaluated mean synthesis quality MOS$_\mu$=2.82 of SVLN is comparably lower than MOS$_\mu$=3.38 for PSY. Stronger voice quality changes are assessed with less good MOS synthesis qualities for both systems. PSY received in general a lower deviation from the true voice quality rating and a higher MOS synthesis quality related to the baseline method SVLN, shown in table 3.

Method	Δ VQ$_\mu$	Δ VQ$_{\sigma^2}$	MOS$_\mu$	MOS$_{\sigma^2}$
PSY	0.8312	0.6858	3.3766	0.9880
SVLN	0.9740	0.5967	2.8182	0.7462

Table 3: *Voice quality (VQ) and MOS sound quality*

6. CONCLUSIONS

The findings presented with the subjective listening test of section 5 suggest that the proposed novel speech analysis and synthesis system PSY is able to analyze an input speech phrase such that different re-synthesized versions carry the perception of different voice quality characteristics. Its assessed synthesis quality received partially very good judgements for minor changes in voice quality. Major voice quality changes are appraised of moderate quality for both the baseline and the proposed method. However, further work is required to render the GMM energy prediction applicable for all cases. Please note that the proposed speech framework will be integrated as system to synthesize singing voices within the ANR project ChaNTeR [3].

[3] ChaNTeR: anasynth.ircam.fr/home/projects/anr-project-chanter/

Acknowledgments

The main author was financed by a CIFRE contract as a former collaboration between the research institute IRCAM and the company Acapela Group. Currently he is financed by a grant from the ANR Project ChaNTeR to enhance the proposed system for singing voices synthesis. He is very grateful for the kind attendance by his supervisor Dr. Axel Röbel and the support from the Acapela Group.

7. REFERENCES

[1] D. G. Childers and C. K. Lee, "Vocal quality factors: analysis, synthesis, and perception." *Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–410, 1991.

[2] J.-S. Liénard and C. Barras, "Fine-grain voice strength estimation from vowel spectral cues," in *14th Annual Conference of the Int. Speech Communication Association (Interspeech ISCA)*, Lyon, France, 2013, pp. 128–132.

[3] J. D. M. Laver, *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980, vol. 31.

[4] J. Kane, S. Scherer, M. Aylett, L.-P. Morency, and C. Gobl, "Speaker and language independent voice quality classification applied to unlabeled corpora of expressive speech," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[5] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 968–981, 2012.

[6] J. P. Cabral and J. Carson-Berndsen, "Towards a better representation of the envelope modulation of aspiration noise," in *Advances in Nonlinear Speech Processing*, ser. Lecture Notes in Computer Science, T. Drugman and T. Dutoit, Eds. Springer Berlin Heidelberg, 2013, vol. 7911, pp. 67–74. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-38847-7_9

[7] D. Vincent, O. Rosec, and T. Chonavel, "A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, 2007, p. 525–528.

[8] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *9th Annual Conference of the Int. Speech Communication Association (Interspeech ISCA)*, Brisbane, Australia, September 2008, pp. 1829–1832.

[9] G. Degottex, P. Lanchantin, A. Röbel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.

[10] X. Serra, *Musical Sound Modeling with Sinusoids plus Noise*. Swets and Zeitlinger, 1997, pp. 91–122. [Online]. Available: files/publications/MSM-1997-Xserra.pdf

[11] G. Fant, "The source filter concept in voice production," *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, vol. 22, no. 1, pp. 021–037, 1981.

[12] R. Maia and Y. Stylianou, "Complex cepstrum factorization for statistical parametric synthesis," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014, pp. 3839–3843.

[13] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, vol. 26, no. 4, pp. 001–013, 1985.

[14] G. Fant, "The lf-model revisited. transformation and frequency domain analysis," *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, vol. 36, no. 2-3, pp. 119–156, 1995.

[15] G. Fant, "The voice source in connected speech," *Speech Communication*, vol. 22, no. 2-3, pp. 125–139, 1997.

[16] S. Huber, A. Röbel, and G. Degottex, "Glottal source shape parameter estimation using phase minimization variants," in *13th Annual Conference of the Int. Speech Communication Association (Interspeech ISCA)*, ser. 1990-9772, Portland, Oregon, USA, 2012, pp. 1644–1647.

[17] S. Huber and A. Röbel, "On the use of voice descriptors for glottal source shape parameter estimation," *Computer, Speech, & Language*, vol. 28, no. 5, pp. 1170 – 1194, 2014.

[18] G. Degottex, A. Röbel, and X. Rodet, "Joint estimate of shape and time-synchronization of a glottal source model by phase flatness," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 5058–5061.

[19] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modelling with unknown model order," *Elsevier, Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343 – 1350, 2007.

[20] C. d'Alessandro, B. Bozkurt, B. Doval, T. Dutoit, N. Henrich, V. Tuan, and N. Sturmel, "Phase-based methods for voice source analysis." in *Advances in Nonlinear Speech Processing*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, vol. 4885, pp. 1–27.

[21] M. Zivanovic and A. Röbel, "Adaptive threshold determination for spectral peak classification," *Computer Music Journal*, vol. 32, no. 2, pp. 57–67, 2008.

[22] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, 1998.

[23] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 4609–4612.

[24] T. Drugman and Y. Stylianou, "Maximum voiced frequency estimation: Exploiting amplitude and phase spectra," *Signal Processing Letters, IEEE*, vol. 21, no. 10, pp. 1230–1234, Oct 2014.

[25] P. Lanchantin and X. Rodet, "Dynamic model selection for spectral voice conversion," in *11th Annual Conference of the Int. Speech Communication Association (Interspeech ISCA)*, Makuhari, Chiba, Japan, 2010, pp. 1720–1723.

[26] P. Lanchantin and X. Rodet, "Objective evaluation of the dynamic model selection method for spectral voice conversion," in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 5132–5135.

[27] B. Doval and C. d'Alessandro, "The spectrum of glottal flow models," Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Orsay), Orsay, Tech. Rep. LIMSI 99-07, 1999.

[28] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[29] N. Henrich, C. d'Alessandro, B. Doval, M. Castellengo, G. Sundin, and D.Ambroise, "Just noticeable differences of open quotient and asymmetry coefficient in singing voice," *Journal of Voice*, vol. 17, no. 4, pp. 481–494, 2003.

[30] G. Degottex, "Glottal source and vocal tract separation," Ph.D. dissertation, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Université de Pierre et Marie Curie (UPMC), Université de Paris 6, Paris, France, 2010.

[31] G. Degottex, A. Röbel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5128–5131.

[32] 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 Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 4630–4633.