Glottal/Supraglottal Source Separation in Fricatives Based on Non-Stationnary Signal Subspace Estimation

Benjamin Elie, Gilles Chardon

Glottal/Supraglottal Source Separation in Fricatives Based on Non-Stationnary Signal Subspace Estimation

Benjamin Elie, and Gilles Chardon

***Abstract*—The X-GLOS (*EXtraction of GLOttal Sources*) method for separating the glottal and the supraglottal sources in speech signals is presented in this article. Unlike other periodic/aperiodic decomposition methods that use stationary models of the signal within frames, X-GLOS considers locally varying instantaneous fundamental frequency. Applications on numerically synthesized fricative signals prove the locally non-stationary model to be more robust to moderate and high jitter values than stationary models. A peak picking selection also allows X-GLOS to be less sensitive to high colored noise levels. The gain of performance, in comparison with the reference existing method, is about a couple of dozens of dB in high noise-to-harmonics ratios. X-GLOS can then be specifically used to study the behavior of the voicing and frication noise sources independently, even at vowel-consonants behavior where the voiced source is less powerful that the frication noise in the recorded mixture speech signal.**

Index Terms: Periodic/aperiodic decomposition, Speech production, Fricative production

I. Introduction

An acoustic speech signal is the sum of the contributions of various acoustic sources, each with their specific properties. Their characteristics and predominance in the resulting speech signal are commonly used as sound classifiers. These sources may be classified into two main categories, : glottal and supraglottal sources. The glottal source mostly includes the oscillation of the vocal folds, responsible for voicing, and the supraglottal sources mostly include the frication noise in fricatives and affricates and burst in stop consonants. In signal processing, these different acoustic sources may be modeled as harmonic and noise sources respectively. In our perspective, we consider the glottal source as the harmonic component of the speech signal, while the supraglottal sources are the residual noise.

Since each individual source gives information that can be used in speech studies and speech analysis, it is important to be able to separate them from the mix speech signal. For instance, the energy ratio between the harmonic and the noise components, named the *Harmonics-to-Noise Ratio* (HNR), is used as an indicator to detect voice pathology [1]–[4].

Numerous methods have been developed in order to separate the noise component of human voice from the periodic component. They can be classified as time domain based [5] or frequency domain based [6]–[9] methods. These previous studies neglected the non-stationarities of the speech signal within the analyzed local frame so that the harmonic component is considered as periodic, hence the periodic/aperiodic denomination. Due to the non-stationary nature of the glottal source, the aperiodic estimate will include contributions of the glottal source, mainly the jitter and the shimmer. This may be an important issue if one wishes to accurately separate the contributions of the glottal and the supraglottal sources.

Another limit of the aforementioned methods is that they are designed for weak noise levels in the analyzed speech signal. Situations where the noise level is expected to be similar or higher than the periodic level are not investigated. Such settings are not uncommon in natural speech, as it is the case during the production of fricatives, for instance. This class of consonants is characterized by the appearance of frication noise due to the generation of turbulence of the air flow downstream of the supraglottal constriction. It comes that at the vowel-fricatives boundaries, the vowel offset is characterized by a gradual decrease of the voiced contributions, whereas the frication noise level increases at the same time. The noise coloration is another difficulty since the resulting noise is colored according to the transfer function of the vocal tract [10]. Being able to separate the voiced component from the frication noise may be helpful for several reasons, such as investigating the spectral characteristics of the noise source alone, e.g. [10], investigating the degree of voicing, for instance in the case of final devoicing [11], or in the case of production strategies [12], [13], tackling problems related to microprosody effects at the vicinity of obstruents [14], improving the acoustic-to-articulatory inversion in the case of fricatives [15], [16], or modifying each individual acoustic source in speech synthesis [17].

Since the paper by Jackson and Shadle [8], which introduced the PSHF (*Pitch Scale Harmonic Filter*) algorithm, it has been considered as the reference method for periodic/aperiodic decomposition. Although it provides an accurate decomposition in many cases, its performance linearly decreased with the noise level, and also with the local non-stationary features, namely the jitter and the shimmer.

A. Our contributions

The main challenge of this paper is then to propose a method that is less sensitive to the noise level and noise coloration, and also less sensitive to the local non-stationary effects of the analyzed speech signal. To tackle these problems, the paper introduces our method, called X-GLOS (*EXtraction of GLOttal Sources*), which proposes a locally non-stationary model of the signal subspace which accounts for the variation of

the instantaneous fundamental frequency inside the analyzed temporal frame. The sensitivity to the frication noise level is reduced by selecting the partials that are actually activated.

After a brief description of the signal models used in this paper in Sec. II, and our pitch detection method in Sec. III, our X-GLOS algorithm is described in Sec IV. Numerical validations are presented in Sec. V in order to quantitatively assess its performance in various cases by using numerically synthesized fricative signals. X-GLOS is compared to the reference method, namely PSHF [8]. Experimental applications on real speech signals are finally presented in Sec. VI.

II. Signal model

In this paper, the speech signal $s(t)$ is decomposed in an harmonic component $s_p(t)$ and a noise component $s_n(t)$:

$$s(t) = s_p(t) + s_n(t) \tag{1}$$

The harmonic component is further decomposed as a sum of M sinusoids

$$s_p(t) = \sum_{m=1}^{M} a_m(t) e^{i\phi_m(t)} \tag{2}$$

where the phases $\phi_m(t)$ are given by

$$\phi_m(t) = 2\pi \int_0^t m f_0(\tau) d\tau + \phi_{0m} \tag{3}$$

and the negative frequencies are implicitly considered in the sum (2) and the following equations.

The fundamental frequency $f_0(t)$ and the amplitudes $a_m(t)$ are slowly varying compared to the period $1/f_0(t)$, and are considered independent to the noise component $s_n(t)$. In practice, it is a strong assumption as the noise component of speech signals may be correlated to the voiced source, resulting in periodic modulations of the frication noise and the glottal noise [18]. For the rest of the paper, we consider a discrete speech signal $s[n]$ sampled at a uniform sampling frequency f_s

In many previous papers (e.g. [7], [8]), although speech signals are by nature not stationary (i.e. the frequency and amplitudes of partials constantly change, as well as the noise characteristics), the analysis is made through short time periods where the signal may be considered as stationary. The discrete speech signal $s[n]$ is then segmented into short frames of length L. To avoid spectral leakage and discontinuities, the signal is windowed by a e.g. a Hann window

$$h[n] = \begin{cases} \frac{1}{2} + \frac{1}{2}\cos\left(2\pi \frac{n}{L}\right) & \text{if } |n| \leq L/2 \\ 0 \text{ if } |n| > L/2 \end{cases}$$

The source separation is applied on frames

$$s_k[n] = s[n] h_k[n],$$

where $h_k[n] = h[n - ka]$ and $a = L/4$. The estimates of the periodic and aperiodic components are then denoted $\tilde{s}_{k,p}$ and $\tilde{s}_{k,n}$ respectively, and are defined as

$$s_k[n] = \tilde{s}_{k,p}[n] + \tilde{s}_{k,n}[n], \tag{4}$$

and the periodic and aperiodic components are estimated by

$$\hat{s}_p[n] = \frac{1}{A} \sum_{k \in \mathrm{Z}} \hat{s}_{k,p}[n] h_k[n] \tag{5}$$

and likewise for $\hat{s}_n[n]$. The constant A depends on the window. Here, $A = 2/3$. From now on, only the signal in the analyzed frame is considered in the description of the signal model, so that the index l, with $l = 0, \ldots, L$, replaces n.

A. Locally stationary sinusoidal model

Let M be the number of activated partials of the voiced source, assuming the signal subspace to be stationary within the local frame k, the periodic component $s_{k,p}$ is then the sum of the contributions of the M sinusoids, hence

$$s_{k,p}[l] = h_k[l] \sum_{m=1}^{M} b_m e^{2\pi j f'_m l}, \tag{6}$$

where $f'_m = \frac{f_m}{f_s}$ and b_m are respectively the normalized frequency and the complex amplitude of the m^{th} sinusoid. Introducing Eq. (6) into Eq. (4), it finally comes

$$s_k[l] = w_k[l] \sum_{m=1}^{M} b_m e^{2\pi j f'_m l} + h_k[l] s_{k,n}[l], \tag{7}$$

or, in a matrix form

$$\mathbf{s}_k = \mathbf{V}_k \mathbf{b}_k + \mathbf{s}_{k,n} = \mathbf{s}_{k,p} + \mathbf{s}_{k,n}, \tag{8}$$

with $\mathbf{s}_k \in \mathbb{R}^{L\times 1}$, $\mathbf{s}_{k,p} \in \mathbb{R}^{L\times 1}$, and $\mathbf{s}_{k,n} \in \mathbb{R}^{L\times 1}$ are the vectors containing the L samples of the windowed mixture, periodic, and aperiodic signals, respectively, $\mathbf{b}_k \in \mathbb{C}^{M\times 1} = [b_1, b_2, \ldots, b_M]^T$ is the vector containing the complex amplitude of the M sinusoids, and where $\mathbf{V}_k \in \mathbb{C}^{L\times M}$ is the base of the signal subspace spanned by the windowed M sinusoids, hence

$$\mathbf{V}_k = \mathbf{H} \begin{bmatrix} 1 & 1 & \ldots & 1 \\ e^{2\pi j f'_1} & e^{2\pi j f'_2} & \ldots & e^{2\pi j f'_M} \\ e^{4\pi j f'_1} & e^{4\pi j f'_2} & \ldots & e^{4\pi j f'_M} \\ \vdots & \vdots & \ddots & \vdots \\ e^{2(L-1)\pi j f'_1} & e^{2(L-1)\pi j f'_2} & \ldots & e^{2(L-1)\pi j f'_M} \end{bmatrix}. \tag{9}$$

where $\mathbf{H}$ is the diagonal matrix containing the samples of the window h.

The principle of the method is then to estimate the periodic component vector $\tilde{\mathbf{s}}_{k,p} = \tilde{\mathbf{V}}_k \tilde{\mathbf{b}}_k$ by first estimating the frequency of the activated partials in order to build $\tilde{\mathbf{V}}_k$, an estimate of the basis of the periodic signal subspace $\mathbf{V}_k$, and then by finding $\tilde{\mathbf{b}}_k$, an estimate of the complex amplitude vector $\mathbf{b}_k$. From the estimate $\tilde{\mathbf{s}}_{k,p}$, it comes the estimate of the aperiodic component $\tilde{\mathbf{s}}_{k,n}$ as

$$\tilde{\mathbf{s}}_{k,n} = \mathbf{s}_k - \tilde{\mathbf{s}}_{k,p}. \tag{10}$$

Note that the stationary model may be useful if one wants the non-stationary components (jitter, shimmer…) to be included into the aperiodic signal. It can be the case for studies about vocal roughness, and voice pathology [1]–[4]. In that

case, the analyzed voice fragments are vowels, namely in high harmonics-to-noise ratio conditions. The method described in our paper is intended to be used to investigate the amount of frication noise embedded in the fricative signal comparatively to the amount of voiced source contributions, namely in low harmonics-to-noise ratio conditions. Consequently, the non-stationary components due to voiced aperiodicity will be considered as voiced source contributions, hence the need to model the voiced signal as non-stationary in the analyzed frame window. The locally non-stationary model is detailed in the next section.

B. Non-stationary sinusoidal model

Now, we assume that the signal subspace is modulated in frequency. In speech, the frequency modulation may be important due to pitch variation and jitter (i.e. the period variation from a glottal cycle to the next) [19]. In that case, the basis matrix $\mathbf{V}_k$ writes

$$\mathbf{V}_k = \mathbf{H} \begin{bmatrix} e^{j\phi_1(0)} & e^{j\phi_2(0)} & \dots & e^{j\phi_M(0)} \\ e^{j\phi_1(1)} & e^{j\phi_2(1)} & \dots & e^{j\phi_M(1)} \\ e^{j\phi_1(2)} & e^{j\phi_2(2)} & \dots & e^{j\phi_M(2)} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\phi_1(L-1)} & e^{j\phi_2(L-1)} & \dots & e^{j\phi_M(L-1)} \end{bmatrix}, \tag{11}$$

where $\phi_m(l) = \frac{1}{f_s}\sum_{i=0}^{l} f_m(i)$, and $f_m(i)$ is the fundamental frequency at the i^{th} sample of the considered frame.

For the rest of the paper, unless explicitly specified, the subscript k will be removed for the sake of clarity.

III. Fundamental frequency estimation

The first step towards the extraction of the signal subspace is the estimation of the fundamental frequency of the voiced components. Many fundamental frequency estimator for speech signals have been proposed in the past (e.g. [8], [20]–[22]). One major difficulty in speech is the high power colored noise that may disturb the pitch detection based on simple Fourier peak picking, or based on autocorrelation, since the noise level is very likely to be higher than harmonic component in certain frequency bands. This is especially true for voiced fricatives.

In order to fix this issue, X-GLOS first selects the potential harmonic candidates: they are selected as peaks of the power spectrum $S(f) = |\hat{s}(f)|^2$ of $s[n]$ that are above a threshold. The estimation of this threshold is crucial and several methods have been tested. A simple choice is to define it as twice the power spectral density of the noise, the latter being estimated from a local median filtering [23]. The first problem with this technique is the fact that the filter order, i.e. the frequency span of the sliding window, had to be set to high value to completely remove the harmonic contribution of the power spectrum. Consequently, many harmonic peaks were likely to be included in the window, so that a relatively weak harmonic could be not detected. Secondly, the threshold had to be set arbitrarily, without any prior knowledge on either the pitch frequency or the noise level. Better results have been obtained by keeping the adaptive filter order and by modifying the threshold by setting it to the value of the power spectrum filtered by a rank filter (or percentile filter) [24] of high rank. We chose to set this rank to 90%, meaning that we keep only the last decile of the power spectrum in the window having a frequency span of $1.25f_0$.

Once the possible partials are identified, the values of the filtered periodogram outside of an interval around the partials are set to zero:

$$S'(f) = S(f)F(f)$$

where $F(f) = 1$ if f is close to an identified partial, and $F(f) = 0$ otherwise. Keeping an interval around each partial is necessary because of the sampling of the frequency axis, as well as possible slight inharmonicity in the signal.

Finally, the estimation of the fundamental frequency uses a cumulative periodogram, similarly to Drugman and Alwan [21]. However, the cumulative periodogram is here computed on the filtered version of the periodogram $S'(mf)$ at multiples of a given frequency:

$$S_c(f) = \sum_{m=1}^{M} S'(mf).$$

As the energy in the spectrum that do not correspond to a partial is eliminated in $S'(f)$, the cumulative periodogram is robust to high levels of noise energy.

The estimation of f_0 from the cumulative periodogram is subject to octave errors. Higher octave errors are possible only if all even partials are not selected as partials, which is unlikely. More probable are lower octave errors, as the energy for a given actual fundamental frequency f is equal to the energy at $f_0/2$, or even slightly lower, if noise has been identified as partials near frequencies $nf_0/2$ for odd n. To deal with such errors, we first select possible f_0 peaks as the peaks higher than a user-defined γ times the higher peak of S_c. The estimated f_0 is chosen as the frequency for which the mean value of the orders of activated partials is the smallest. In the case of an octave error, this mean value is double for the lower octave.

IV. Source separation

Once the fundamental frequency has been estimated, the next steps consist in, firstly, precisely estimating the frequencies of the activated partials in order to build the basis of the voiced signal subspace $\mathbf{V}_k$ in Eq. (9), and secondly, in estimating the complex amplitudes $\mathbf{b}_k$.

A. Estimation of the periodic signal subspace

For every frames where voicing activity has been detected, i.e. $f_0 > 0$, the frequency of the activated partials are estimated using the *Quadratically Interpolated FFT* (QIFFT) method [25]. This technique consists in estimating the sinusoidal parameters analytically from a second-order polynomial model derived from the log amplitude of the FFT bins corresponding to the spectral peak and its two neighbors. Here, this method is used only to estimate the sinusoid frequency: it is then the frequency of the maximum of the polynomial, i.e. $f_m = -\frac{p_{2,m}}{2p_{1,m}}$ where $p_{1,m}$ and $p_{2,m}$ are the first and

second order coefficients of the polynomial associated to the m^{th} partial. To ensure an unbiased frequency estimation, the QIFFT should be used with Gaussian windows, since the log-magnitude of such windows is actually a second order polynomial.

Considering the voiced source as periodic in first approximation, the spectral peaks should be located at frequencies that are multiples of the fundamental frequency. If this is not the case, namely if the m^{th} spectral peak is located outside a arbitrarily defined interval around mf_0, the partial is considered as not activated. In this paper, the interval is defined as ± 5 Hz, that is, the m^{th} spectral peak is searched in the frequency range $f_{m-1}+f_0 \pm 5$ Hz. The reference harmonic f_{m-1} is updated in order to avoid error spreading in high order spectral peaks. In high noise level conditions, and more specifically in the case of the appearance of frication noise at a high level, the voiced contributions are very small ahead of the noise in the high-frequency range. This is due to the fact that the voiced source has a low-pass profile, characterized by a relatively strong spectral slope [26], while the frication noise contains high level components in the mid- and the high-frequency domains [10]. The critical frequency above which voiced contributions are considered as negligible is sometimes referred as the *Maximum Voiced Frequency* (MVF) [21]. In order to avoid including noise components into the voiced signal subspace, we propose to make a rough estimation of the MVF and to consider all components above the MVF estimate as noise components. The MVF is taken as the last activated partial before a at least a certain number of partials, denoted by N_z, are not detected in a row.

Once the frequencies of every harmonic candidates have been estimated, the basis of the voiced subspace is computed following Eq. (9) for the stationary model, and following Eq. (11) for the locally non-stationary model. In the second case, the instantaneous frequency is computed by using a zero-crossing estimation of the band-pass filtered version of the analyzed frame, where the bandwidth of the filter is 10 Hz, centered around the rough pitch estimate. A first-order linear interpolation is then performed to compute the instantaneous frequency between two successive zero-crossing instants.

B. Estimation of the complex amplitudes

The problem of the amplitude estimation of sinusoids in the presence of noise has been tackled in the past, e.g. in [27]. The classic method is the *Least Square* (LS) estimation. Considering our case, the LS estimate of the complex amplitudes $\mathbf{b}_k$ is

$$\mathbf{b}_k = \mathbf{V}^{\dagger}\mathbf{s}_k, \tag{12}$$

where $(.)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse. In this approach, the noise is considered white and homoscedastic.

Several methods have been proposed, such as the *Weighted Least Square* (WLS) [28], the *Capon* [29], and the *Amplitude and Phase Estimation* (APES) [30], in order to outperform the LS method by accounting for noise correlations. This is an important issue in periodic/aperiodic decomposition, especially in the presence of high noise level. Yet, in our case, none of the aforementioned methods have been shown to significantly modify or improve the results of the separation. Consequently, we chose to keep the LS method because of its comparatively low computational cost.

V. Numerical validation

A. Synthetic signals

Numeric voiced signals and frication noise signals were generated separately by using a *Transmission Line Circuit Analog*-based speech synthesizer [31]. Synthetic signals correspond to simulations of French fricatives at two places of articulations, namely the alveolar voiced/unvoiced pair /z,s/ and the post-alveolar pair /ʒ,ʃ/. To simulate these fricatives, the voice synthesizer needs to be fed with area functions and glottal opening area waveforms. Our simulations used two area functions extracted from 3D static MRI of a French male native speaker of 35 years old at the time of acquisition, corresponding to the alveolar and the post-alveolar French fricatives in the vocalic context /a/. For voiced signals, the glottal opening area function were first generated with a parametric model for each fundamental frequency and jitter value. The resulting glottal opening area waveform is then used to feed the synthesizer for both area functions, with the frication noise generator disabled in order to get a purely voiced signal in output. Then, the frication noise signal is obtained by setting a constant glottal area opening in input of the synthesizer with the frication noise generator enabled. Doing so ensures the synthetic voiced and frication noise signals to be as similar to those of real fricatives as possible.

The jitter effect was simulated similarly to Jackson and Shadle [8], by modifying the nominal period value T_0 by a normal random distribution of zero mean and variance equal to the jitter:

$$\tau_i = T_0 \left(1 + J\frac{r_i\sqrt{\pi}}{2}\right), \tag{13}$$

where τ_i is the value of the considered period, r_i is a random variable from the normal distribution of zero mean and unit standard deviation, and J is the jitter value.

Then, once the voiced and the noise signals, respectively denoted $\hat{s}_p$ and $\hat{s}_n$, are simulated, they were normalized by their respective norm, and their energy were scaled by a coefficient α in order to simulate various theoretical voicing quotients:

$$s_{mix} = (1-\alpha)\frac{\hat{s}_p}{||\hat{s}_p||_2} + \alpha\frac{\hat{s}_n}{||\hat{s}_n||_2} = \bar{s}_p + \bar{s}_n, \tag{14}$$

where s_{mix} is the input mix signal, and $\bar{s}_p$ and $\bar{s}_n$ are the target voiced and noise signals to be estimated, and where $||.||_2$ denotes the ℓ_2-norm. The scaling factor α is defined according to the desired voicing quotient VQ:

$$\alpha = 1 - \sqrt{VQ}.$$

For the simulations, the nominal fundamental frequency is varied from 120 to 200 Hz by a 20 Hz incremental step, the jitter value from 0 to 3% with an incremental step of 0.5%, and the voicing quotient from 0 to 100% with an incremental step of 5%.

Note that the synthetic signals follows the model in Eq. (1), namely glottal and supraglottal sources are additive, which is not necessary the case in natural speech, as noise amplitude modulation by the glottal pulses may occur. However, it is not predominant, especially in high noise level conditions. Decomposition methods are commonly tested with this additive model, and, consequently, the reader should be aware that the presented performances are biased. In practice, they are likely to be slightly less accurate.

B. Performance indicators

In order to be able to compare X-GLOS with another method, namely the PSHF [8], the *Signal-to-Error Ratio* (SER) defined in [8] is used. It is defined both for the voiced and the unvoiced components as

$$\eta_p = 10\log_{10}\left(\frac{||\bar{s}_n||_2^2}{||e||_2^2}\right) \quad \text{(dB)} \tag{15}$$

$$\eta_n = 10\log_{10}\left(\frac{||\bar{s}_p||_2^2}{||e||_2^2}\right) \quad \text{(dB)}, \tag{16}$$

where $e = \tilde{s}_p - \bar{s}_p$ is the estimation error.

C. Results

1) Effect of the fundamental frequency: The performance of the PSHF algorithm [8] has been shown to vary little with the fundamental frequency thanks to the pitch-synchronous technique. This section verifies if this property also applies to X-GLOS. Fig. 1 shows the performance for signals of various fundamental frequencies and without jitter. In the studied range of f_0, i.e. from 120 to 200 Hz, the performance of X-GLOS shows small variations with the pitch value, both for the aperiodic and the periodic estimates. Differences from 120 Hz to 200 Hz are no more than 3 dB, both for high and low voicing quotient conditions, and for alveolar and post-alveolar places of articulation. These variations are in the same order of magnitudes than those observed in the performance of the PSHF [8]. Note that the performance is better with higher pitch values.

For the sake of clarity, since there is no significant modification of the performance with the pitch value, and especially no more than the baseline to which X-GLOS is compared in the next section, next results are presented for a nominal pitch value of $f_0 = 120$ Hz throughout the rest of the paper.

2) Comparison with PSHF: This section compares the performance of X-GLOS to the performance of PSHF [8]. Fig. 2 shows the performance of the tested methods for a 120 Hz-pitch signal with no jitter, for a post-alveolar (top) and an alveolar (bottom) fricative, and as a function of the voicing quotient VQ. The values of the performance indicators η_n and η_p obtained with PSHF are similar to those shown in the original paper (see Fig. 3 in [8]), namely η_n is slightly less than 0 in low voicing conditions and around 25 dB for high voicing quotient conditions, and η_p is constantly around 5 dB. The figure shows that X-GLOS constantly outperforms PSHF, both in high and low voicing conditions. Without MVF filtering (solid line, box marker), the gain of performance with X-GLOS in comparison with PSHF is between 2.5 and 5 dB

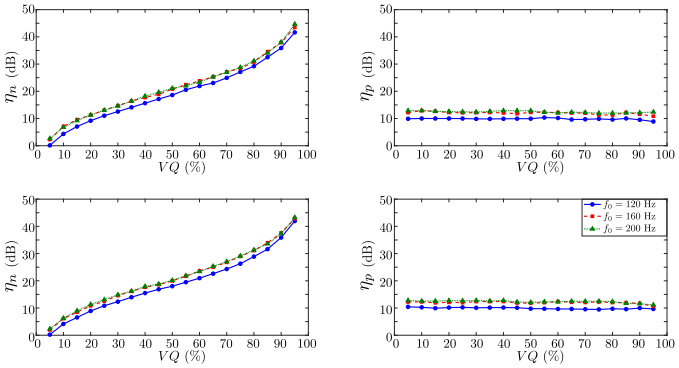

Figure 1: Aperiodic (left) and periodic (right) performance of the proposed method with no MVF selection for different pitch values, $f_0 = 120$ Hz (solid line, circle marker), $f_0 = 160$ Hz (dashed, square marker), and $f_0 = 200$ Hz (dotted line, Δ marker). Top is the post-alveolar fricative simulation, bottom is the alveolar fricative simulation.

for both the periodic and the aperiodic estimates. Changing the place of articulation has no significant influence on the performances.

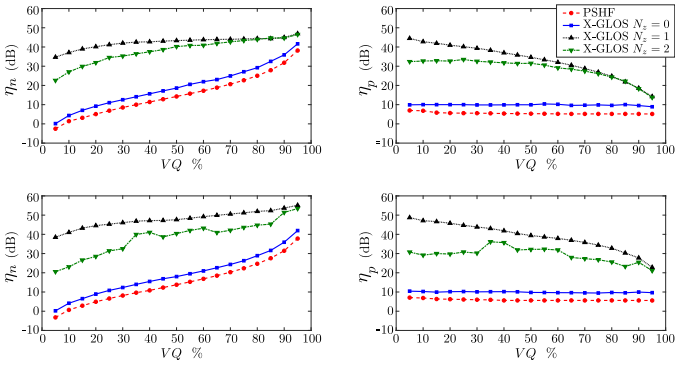

Figure 2: Aperiodic (left) and periodic (right) performance of X-GLOS with no MVF selection (solid line, box marker), with first zero MVF selection (dotted, Δ marker), and with two first zeros MVF selection (dot-dashed line, ∇ marker). The performance of PSHF is plotted as the dashed line, circle marker. Top is the post-alveolar fricative simulation, bottom is the alveolar fricative simulation.

When a MVF filtering is applied, the performance of X-GLOS is significantly enhanced. Basically, the method becomes less sensitive to the noise level: between the almost no noise condition ($VQ = 95\%$) and the quasi-unvoiced condition ($VQ = 5\%$), the SER of the aperiodic estimate η_n is lowered by 24 dB with a first-zero MVF filtering, and by only 12 dB for a second-zero MVF filtering, whereas it falls by 41 dB without MVF filtering. For the periodic estimate η_p, applying the MVF filtering improves by 20 dB for a first-zero MVF filtering and by 30 to 40 dB with a second-zero filtering in low voicing conditions, in comparison to the no-MVF filtering situation.

3) Effect of the jitter: This section discusses the performance of X-GLOS for various jitter values. It also introduces the non-stationary model described in Sec. II-B and compares

its performance with the stationary model. In Fig. 3, the performances are displayed for several values of the jitter (0.5%, 1%, 1.5%, and 3%). The aperiodic estimate performance η_n is systematically better with the non-stationary model than with the stationary model. It is also the case for the periodic estimate performance η_p. Interestingly, whereas the performance of the stationary model significantly decreases as the jitter value increases, it is much less significant for the non-stationary model. The aperiodic performance indicator η_n loses around 4 dB when the jitter goes from 0.5% to 3% with the non-stationary model, while it decreases of 7 dB with the stationary model. Consequently, using the non-stationary model enables the method to be less sensitive to the local pitch fluctuation in natural speech.

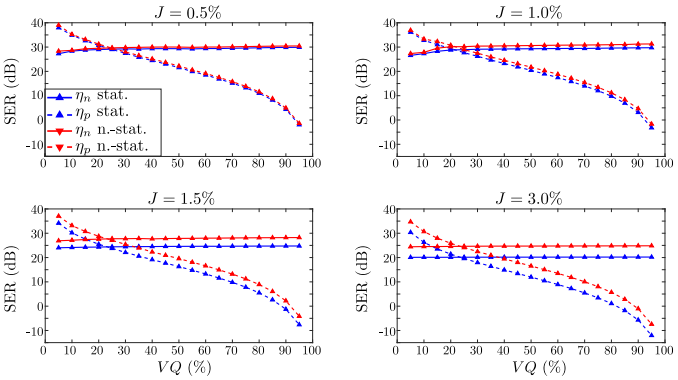

Figure 3: Aperiodic (solid line) and periodic (dashed line) performance of X-GLOS with 2-zeros MVF selection, with a stationary model (Δ marker), and non-stationary model (∇ marker), for various jitter values (0.5%, 1%, 1.5%, and 3%). For the sake of clarity, only the post-alveolar fricative simulation is displayed.

VI. Applications on real speech

For all of the examples shown in this section, data were acquired in an acoustically designed room to reduce background noise, at a sampling rate of 16 kHz. The paper shows examples of utterances of three different speakers, labeled 01F35, 02F24, and 03M33. They were two female speakers, 01F35 and 02F24, and one male speaker, 03M33, and are respectively 35, 24, and 33 years old. 01F35 and 03M33 are French native speaker, while 02F24 is a Basque native speaker with an advanced fluency level of French. All of these speakers reported no speech or hearing impairments. The code and the experimental results are available at `http://gilleschardon.fr/xglos/`.

A. Vowel-consonant-vowel pseudowords

The first experiment consists in pseudowords uttered by the two female speakers, 01F35 and 02F24. Pseudowords are in the form vowel-fricative-vowel (VFV), where the vowel is /ɑ/, and the fricatives are chosen among /ʃ,s,ʒ,z/. In this paper, pseudowords are presented by place of articulation, namely post-alveolar (pseudowords /ɑʃɑ/-/ɑʒɑ/, in Fig. 4), and alveolar (pseudowords /ɑsɑ/-/ɑzɑ/, in Figs. 5 and 6). The post-alveolar example (displayed in Fig. 4) is uttered by 01F35, while the alveolar example (displayed in Fig. 5) is uttered by 02F24.

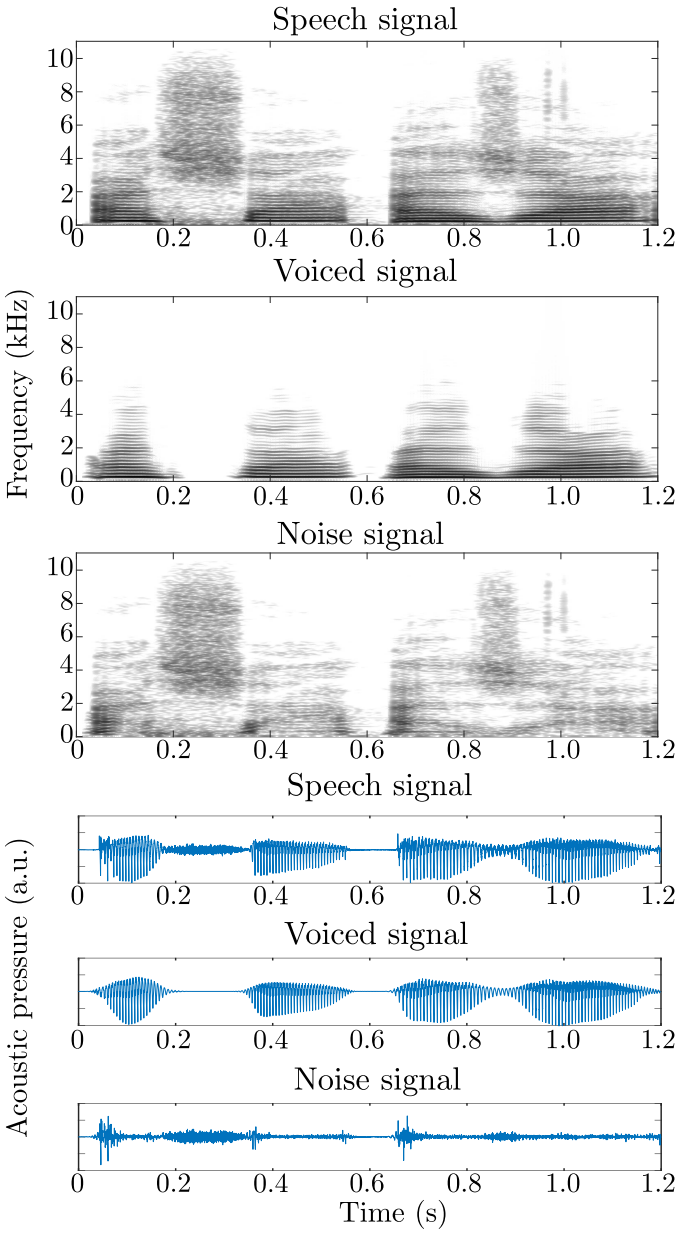

Figure 4: Narrow-band spectrogram (left) and time waveform (right) of the original speech signal (top), the voiced contributions (middle), and the noise estimation (bottom). The original utterance is the two pseudowords "asha-azha" (/ɑʃɑ/-/ɑʒɑ/) uttered by 01F35.

In both examples, i.e. for post-alveolar and alveolar fricatives, displayed in Figs. 4 and 5, X-GLOS proves to efficiently separate the voiced component from the turbulent noise components. Here, we are interested in the vowel-fricative boundaries and in the fricative segments. In any case, the voiceless fricatives segments have been correctly identified, so that only the turbulence noise appears in the separation. In the transitory segments between vowels and fricatives, the separation clearly enables the appearance/disappearance moments of the turbulent noise and of the vocal folds oscillation to

be observed and identified. For voiced fricatives, as expected, the separation shows two distinct frequency zones: the low frequency range where voiced components are predominant, and the high frequency where only turbulent noise is present.

The zoomed-in version of the separated signals in Fig. 6 shows the modifications of the voiced and noise components during the alveolar voiced fricative /z/ in the utterance /ɑzɑ/. The voiced component "loses" its high order harmonic components leading to an almost sinusoidal waveform. The perturbations visible in the speech signal waveform (top figure) have been removed, and constitute the noise signal, shown in the bottom plot. The noise components contain modulations by the voiced components: the noise amplitude is larger in the decreasing phase of the voiced component signal. The presence of noise source modulations is in agreement with previous observations and studies (e.g. in [18]).

B. Phrase-level utterances

The chosen sentence for the example of a phrase-level utterance is the French "L'ours nagea de banquise en banquise" (/luʁs.na.ʒa.də.bɑ̃.ki.zɑ̃.bɑ̃.kiz/), meaning "The bear swam from ice floe to ice floe", uttered by 03M33, a 33 years old French native male speaker. It has been chosen because it contains several natural classes that are of interest, namely fricatives (voiced /ʒ,z/ and voiceless /s/) and stops (voiced /d,b/ and voiceless /k/). The result of the decomposition is shown in Fig. 7.

Similarly to the pseudowords shown in Sec. VI-A, X-GLOS proves to be efficient to separate the contribution of the voiced source from the mixture signal. The voiceless fricative /s/, at $t = 0.50$s, is detected as voiceless, hence zero values in the voiced signal. The decomposition has also detected a low-frequency periodicity at the /ʁ-s/ boundary between $t = 0.35$s and $t = 0.40$s. For voiced fricatives, namely the post-alveolar /ʒ/ at $t \in [0.85, 0.90]$ s, the two alveolar /z/ at $t \in [1.65, 1.75]$ s, and $t \in [2.30, 2.45]$ s, only the low frequency components corresponding to the voiced components appear in the voiced estimate, while the frication noise, at higher frequencies, appears only in the noise estimate. For the bursts of the two velar voiceless stops /k/ at $t \in [1.40, 1.48]$ s and $t \in [2.05, 2.14]$ s, even though a very weak low frequency periodicity has been found, the MVF filtering removed most of the higher frequency components, resulting in a quasi-null estimated periodic signals during these instants.

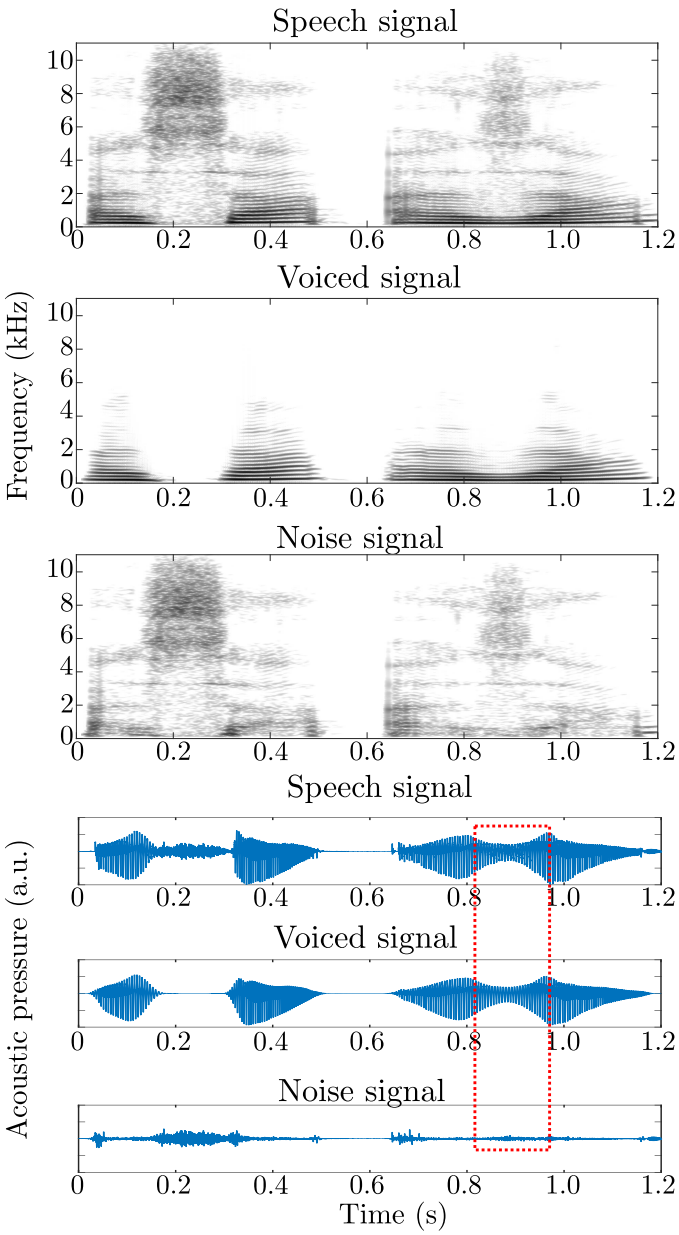

Figure 5: Narrow-band spectrogram (left) and time waveform (right) of the original speech signal (top), the voiced contributions (middle), and the noise estimation (bottom). The original utterance is the two pseudowords "asa-aza" (/ɑsɑ/-/ɑzɑ/) uttered by 02F24. The box in the right column corresponds to the zoom-in region displayed in Fig. 6.

VII. Conclusion

This paper has presented a method, called X-GLOS, for separating the contributions of the supraglottal sources from the contributions of the glottal source in the audio speech signal. A non-stationary model of the glottal source has been used in order to improve the robustness of the method to pitch variations within the analysis window. Additionally, a Maximum Voiced Frequency (MVF) filtering is applied to avoid high frequency noise components to be included into the glottal source estimate.

Tests on synthetic additive signals have shown that these considerations significantly improve the robustness of the decomposition to the frication noise level and coloring, and also the robustness to the jitter. The method has been proved to be less sensitive to the noise level that existing methods, such as PSHF [8]. Applying MVF filtering theoretically improves the performance by up to 20 dB with a first-order MVF filtering and up to 40 db with a second-order MVF filtering. In normal jitter value condition, namely when the jitter is between 0.5 and 1.0 %, the gain on the voiced estimate is around 4 dB in comparison with the stationary model.

The significant decrease of the decomposition performance with high frication noise level and high jitter were the main

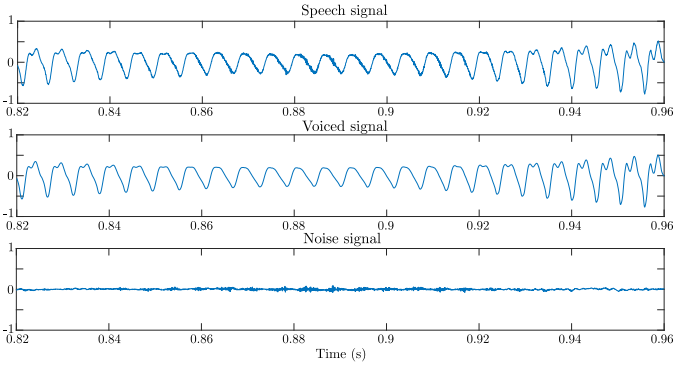

Figure 6: Zoom-in version of the time waveform in Fig. 5. from top to bottom: original speech signal, the voiced contributions, and the noise estimation.

limitations of the previous methods, such as PSHF [8], or PAP [7]. Experiments on real speech signals have shown that X-GLOS can be used to precisely investigate the behavior of the glottal source contributions at the vowel offsets preceding a consonant, either voiced or voiceless, namely when it is drawn into the noise contributions in the mixture signal.

REFERENCES

[1] Eiji Yumoto, Wilbur J Gould, and Thomas Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[2] Hideki Kasuya, Shigeki Ogawa, Kazuhiko Mashima, and Satoshi Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1329–1334, 1986.

[3] Dirk Michaelis, Tino Gramss, and Hans Werner Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[4] Kumara Shama, Niranjan U Cholayya, et al., "Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 50–50, 2007.

[5] Yingyong Qi, Bernd Weinberg, Ning Bi, and Wolfgang J Hess, "Minimizing the effect of period determination on the computation of amplitude perturbation in voice," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2525–2532, 1995.

[6] Guus de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[7] B Yegnanarayana, Christophe d'Alessandro, and Vassilis Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 1, pp. 1–11, 1998.

[8] P. J. B. Jackson and C Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence noise components in speech," *IEEE Trans. Speech Audio Process.*, vol. 9(7), pp. 713–726, 2001.

[9] Nicolas Sturmel, *Analyse de la qualité vocale appliquée à la parole expressive*, Ph.D. thesis, Université Paris Sud-Paris XI, 2011.

[10] C. H. Shadle, *Articulatory-Acoustic relationships in fricative consonants*, Kluwer academic publishers, Dordrecht, 1990.

[11] C. M. R. Pinho, L. M. T. Jesus, and A. Barney, "Weak voicing in fricative production," *Journal of Phonetics*, vol. 40, pp. 625–638, 2012.

[12] Benjamin Elie and Yves Laprie, "Acoustic impact of the gradual glottal abduction degree on the production of fricatives: A numerical study," *The Journal of the Acoustical Society of America*, vol. 142, no. 3, pp. 1303–1317, 2017.

[13] Benjamin Elie and Yves Laprie, "Glottal opening and strategies of production of fricatives," in *Interspeech 2017*, 2017.

[14] James P. Kirby and D. Robert Ladd, "Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2400–2411, 2016.

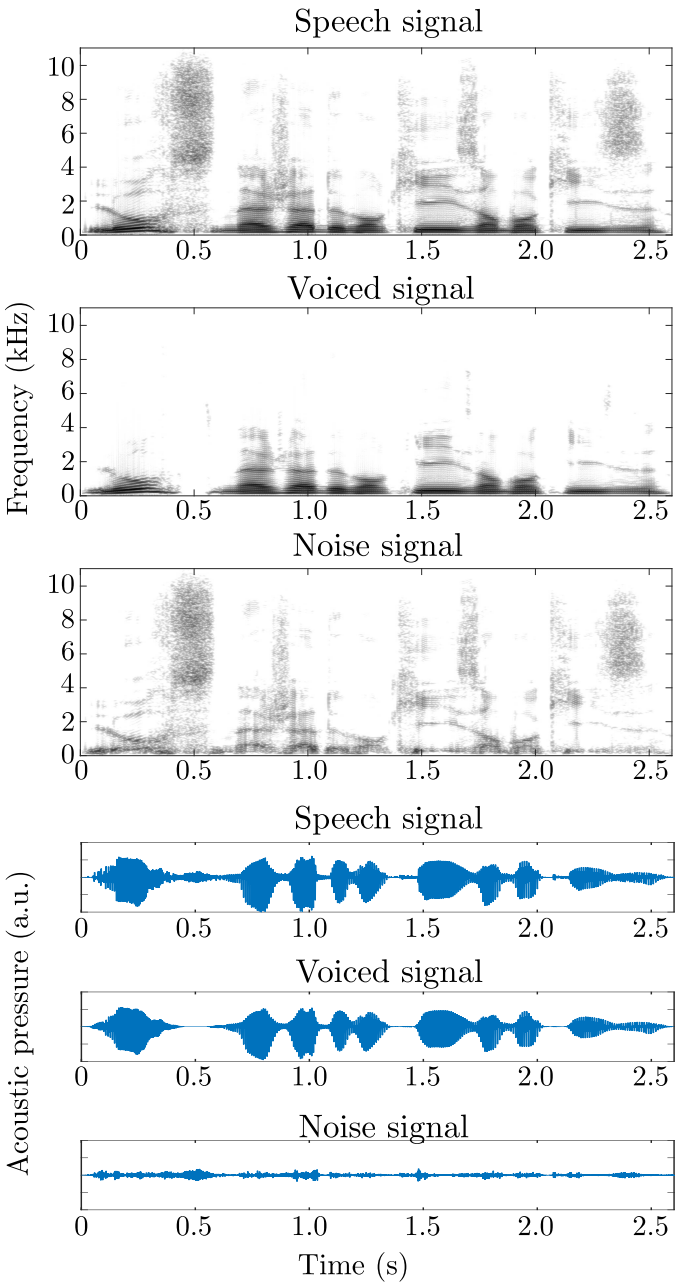

Figure 7: Narrow-band spectrogram and time waveform of the original speech signal (top), the voiced contributions (middle), and the noise estimation (bottom). The original utterance is the French "L'ours nagea de banquise en banquise" (/luʁs.na.ʒa.də.bɑ̃.ki.zɑ̃.bɑ̃.kiz/) uttered by 03M33.

[15] E. L. Riegelsberger, *The acoustic-to-articulatory mapping of voiced and fricated speech*, Ph.D. thesis, Ohio state university, 1997.

[16] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model," *J. Acoust. Soc. Am.*, vol. 129(4), pp. 2144–2162, 2011.

[17] Daniel Erro, Inaki Sainz, Eva Navas, and Inma Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 2, pp. 184–194, 2014.

[18] Anna Barney and Philip J Jackson, "Analysis of frication noise modulation from a physical model," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3578–3578, 2008.

[19] Philip Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 344–353, 1963.

[20] Alain de Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 11(4)1, pp. 1917–1930, 2002.

[21] Thomas Drugman and Abeer Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[22] Meysam Asgari and Izhak Shafran, "Improving the accuracy and the robustness of harmonic model for pitch estimation.," in *Interspeech*, 2013, pp. 1936–1940.

[23] Benjamin Elie and Gilles Chardon, "Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives," in *Proceedings of the 22th International Congress on Acoustics*, 2016.

[24] RM Hodgson, DG Bailey, MJ Naylor, ALM Ng, and SJ McNeill, "Properties, implementations and applications of rank filters," *Image and Vision Computing*, vol. 3, no. 1, pp. 3 – 14, 1985.

[25] M. Abe and J. O. Smith III, "Am/fm rate estimation for time-varying sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Philadelphia, PA*, Mar. 2005, vol. 3, pp. 201–204.

[26] J Sundberg and J Gauffin, "Waveform and spectrum of the glottal voice source," *Frontiers of speech communication research*, pp. 301–320, 1979.

[27] Petre Stoica, Hongbin Li, and Jian Li, "Amplitude estimation of sinusoidal signals: survey, new results, and an application," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 338–352, 2000.

[28] Torsten Söderström and Petre Stoica, *System identification*, Prentice-Hall, Inc., 1988.

[29] Jack Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[30] Jian Li and Petre Stoica, "An adaptive filtering approach to spectral estimation and sar imaging," *IEEE Transactions on Signal Processing*, vol. 44, no. 6, pp. 1469–1484, 1996.

[31] Benjamin Elie and Yves Laprie, "Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink," *Speech Communication*, vol. 82, pp. 85–96, 2016.