Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives

Benjamin Elie, Gilles Chardon

ICA 2016 BUENOS AIRES
Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century...

PROCEEDINGS of the 22nd International Congress on Acoustics

Speech Communication: Paper ICA2016-722

Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives

Benjamin Elie[(a)], Gilles Chardon[(b)]

[(a)]Loria, Inria/CNRS/Université de Lorraine, Nancy, France, benjamin.elie@loria.fr
[(b)]Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec, CNRS, Univ Paris Sud, Université Paris-Saclay, Gif-sur-Yvette, France, gilles.chardon@centralesupelec.fr

August 29, 2016

Abstract

This study presents a method for separating periodic and aperiodic components embedded in speech signals. The fundamental frequency is first estimated from a frequency based technique using a whitened cumulative periodogram. A simple partial detector is used to avoid octave errors. The pitch detection is robust with respect to high level of colored noise. The periodic component is then estimated via the projection of the signal on the subspace spanned by the harmonics. The aperiodic component is obtained by subtracting the periodic component to the analyzed signal. Numerical validations on synthetic signals show that the presented method successfully separate the periodic and aperiodic components of simulated voice segments, even in very complicated case, such as voiced fricatives, which exhibit low and frequency-dependent harmonics-to-noise ratio. Applications on real speech signals highlight the interest of the technique to quantitatively estimate speech features such as harmonics-to-noise ratio, or voicing degree, as a function of time.

Keywords: Fricatives, perodic/aperiodic separation

Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives

1 Introduction

The acoustic signal from human voice contains aperiodic components that come from various sources, e.g. frication noise, breathiness, jitter, shimmer, etc. This leads to a commonly accepted representation of the speech signal as a sum of two main components, called periodic (harmonic, voiced, or tonal) and aperiodic (or noise) components. The relative level of the aperiodic component may be an indicator to detect voice pathologies [1, 13]. 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 [8] or frequency domain based [1, 12, 13] methods.

The aforementioned methods are usually validated and used in weak noise levels. Situations where the noise level is expected to be similar or higher than the periodic level are not investigated. For instance, this situation may occur during the production of fricatives. Indeed, the production of this class of consonant consists in making a narrow supraglottal constriction so that the air flow becomes turbulent downstream the constriction. The turbulence generates frication noise, which is the main acoustic feature of fricative consonants. Voiced fricatives are fricatives which are produced by both generating frication noise, and by making the vocal folds oscillate to generate a voiced source. In this case, the resulting speech signal contains an aperiodic component at high energy level in addition to the voiced component due to the oscillation of the vocal folds. Besides, since the noise source is located inside the vocal tract, downstream of the supraglottal constriction [11], the resulting noise is colored according to the transfer function of the vocal tract. 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. [11], to investigate the degree of voicing, for instance in the case of final devoicing [7], to improve the acoustic-to-articulatory inversion in the case of fricatives [6, 10], and also for modifying each individual acoustic source in speech synthesis [4].

The main challenge in f_0 estimation and periodic/noise separation is the presence of colored noise. In most signal processing methods, the noise is supposed white, or at least of known power spectral density. The proposed method is robust to unknown color noise, through a preliminary estimation of the noise level. Moreover, a precise modeling of the periodic component makes it possible to apply the method on short frames, of length of the order of 6 periods.

The method, detailed in Sec. 2, is numerically validated in Sec. 3 on simulated voiced fricative signals. In Sec. 4, applications on real speech signals illustrate the interest of the method to investigate the production of fricatives by separating the frication noise from the voiced signal.

2 Estimation of periodic components

To estimate the periodic component s_p and the noisy component s_n of a speech signal (or a more general signal) s:

$$s(t) = s_p(t) + s_n(t),$$

the first step is to consider the sound on short periods of time where it can be considered stationary (i.e. constant frequency and partial amplitudes, as well as constant noise statistics). To this end, we use a window h with compact support of length L. A usual choice for h is the Hann window

$$h(t) = \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi t}{L}\right)$$

for $t \in [-L/2, L/2]$ and 0 otherwise. The periodic and noisy components estimation is carried out on extracts of the signals indexed by k :

$$s_k(t) = s(t)h(t - ka)$$

yielding the estimations $\tilde{s}_{k,p}$ and $\tilde{s}_{k,n}$

With $a = L/4$, we have the reconstruction formula

$$s(t) = A\sum_k s_k(t)h(t - ka)$$

with $A = 2/3$, and the periodic and noisy components are estimated by

$$\tilde{s}_p(t) = A\sum_k \tilde{s}_{k,p}(t)h(t - ka)$$

$$\tilde{s}_n(t) = A\sum_k \tilde{s}_{k,n}(t)h(t - ka).$$

2.1 f_0 estimation in presence of colored noise

The first step towards a separation of the periodic and aperiodic components of a signal is the estimation of the fundamental frequency of the periodic component. The periodic component after windowing writes

$$s_p(t) = \sum_{m=1}^{M} h(t)\cos\left(2\pi m f_0 t + \phi_m\right).$$

Ideally, this estimation should be robust to a high power colored noise, as can be the case in speech signals. In this case, simple Fourier or autocorrelation based methods may fail, as the noise level in a certain frequency band could be higher than periodic component level and perturb the f_0 estimation.

To reduce the influence of the noise, the first step of the estimation is the identification of possible partials of the periodic component. Possible partials are identified as peaks of the power spectrum $S(\omega) = |\hat{s}(\omega)|^2$ of $s(t)$ (the subscript n is left out for clarity) with energy larger

than the power of noise in a frequency band around the peak multiplied by a user-defined factor.

To estimate the power spectral density of the noise, the periodogram of the complete signal is passed through a median filter. The output $S_m(\omega)$ is equal to the median of the power spectrum on an interval of width $\Delta\omega$:

$$S_m(\omega) = \underset{\nu\in[\omega-\Delta/2,\omega+\Delta/2]}{\text{median}} S(\nu).$$

The output of the median filter is a robust estimator of the power of the noise in a given frequency band, as the peaks corresponding to partials can be considered as outliers (we refer the reader to [9] for more details on this method). The output of the median filter is therefore an good estimator of the PSD of the noise $P(\omega)$:

$$P(\omega) \approx 1.4836\ S_m(\omega).$$

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

$$S'(\omega) = S(\omega)F(\omega)$$

where $F(\omega) = 1$ if ω is close to an identified partial, and $F(\omega) = 0$ 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. A cumulative periodogram is then computed, by summing the values of the periodogram at multiples of a given frequency:

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

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.

The end result is an f_0 estimation that is robust to

- colored noise, by estimating the PSD of the noise, and selecting possible partials as peaks emerging from the noise level
- octave errors, by comparing the orders of the activated partials from possible f_0 peaks.

2.2 Separation of periodic and aperiodic components

Once the fundamental frequency is estimated, the periodic component is obtained by orthogonally projecting the signal on the space spanned by the sinusoids with frequencies multiple of f_0 (multiplied by the window h). To avoid capturing too much noise in the periodic component, only the partials that are identified as activated in the previous section are considered.

If a proper windowing is applied (e.g. Hann window), and the length of the window is sufficiently large compared to the period, the partials can be considered orthogonal, and the computation of the projection is simplified, as the sinusoids form a basis that is approximately orthogonal.

3 Numerical validation

3.1 Test signals

Numerical validation consists in applying the method presented in Sec. 2 to synthetic signals corresponding to simulated voiced fricatives with various voicing quotient (VQ) and fundamental frequency f_0. The voicing quotient is defined as the proportion of the energy of the periodic component in the speech signal, and is expressed in percent. The voicing quotient VQ is then

$$VQ(\%) = 100 \times \frac{||s_p||_2^2}{||s_p + s_n||_2^2}. \tag{1}$$

The simulated voiced fricatives correspond to the three different articulation place of fricative production in French: labiodental (/v,f/), alveolar (/s,z/), and palato-alveolar (/ʃ,ʒ/). They have been simulated using the time-domain continuous speech synthesizer described in [2,3]. Area functions are derived from static 3D MRI of the vocal tract of a native male French speaker, who was 35 years old at the time of acquisition. The voiced component $s_p(t)$ is obtained by imposing the glottal opening area function in input and by disabling the generation of the frication noise. The aperiodic component $s_n(t)$ is obtained by imposing an arbitrary constant glottal chink area in input, set to 0.4 cm^2, without the periodic perturbations generated by the oscillation of the vocal folds. The mix signal $s(t)$ is then computed as the weighted superimposition of the periodic and aperiodic components, such as

$$s(t) = (1-\alpha)s_p(t) + \alpha s_n(t),, \tag{2}$$

where α is the weighting coefficient used to modify the global noise level.

In the simulations, α linearly varies from 0 to 1 with an incremental step of 0.05. For each of the 21 values of α, 3 different fundamental frequencies f_0 (140, 210 and 340 Hz) have been tested in order to assess the robustness of the method to the number of considered periods.

Fig. 1 shows an example of periodic/aperiodic separation of a simulated palato-alveolar fricative /ʒ/ having a voicing quotient $VQ = 50\%$, i.e. the energy of the periodic component is equal to the energy of the aperiodic component. The results show that the estimated signals are very similar to the simulated signals.

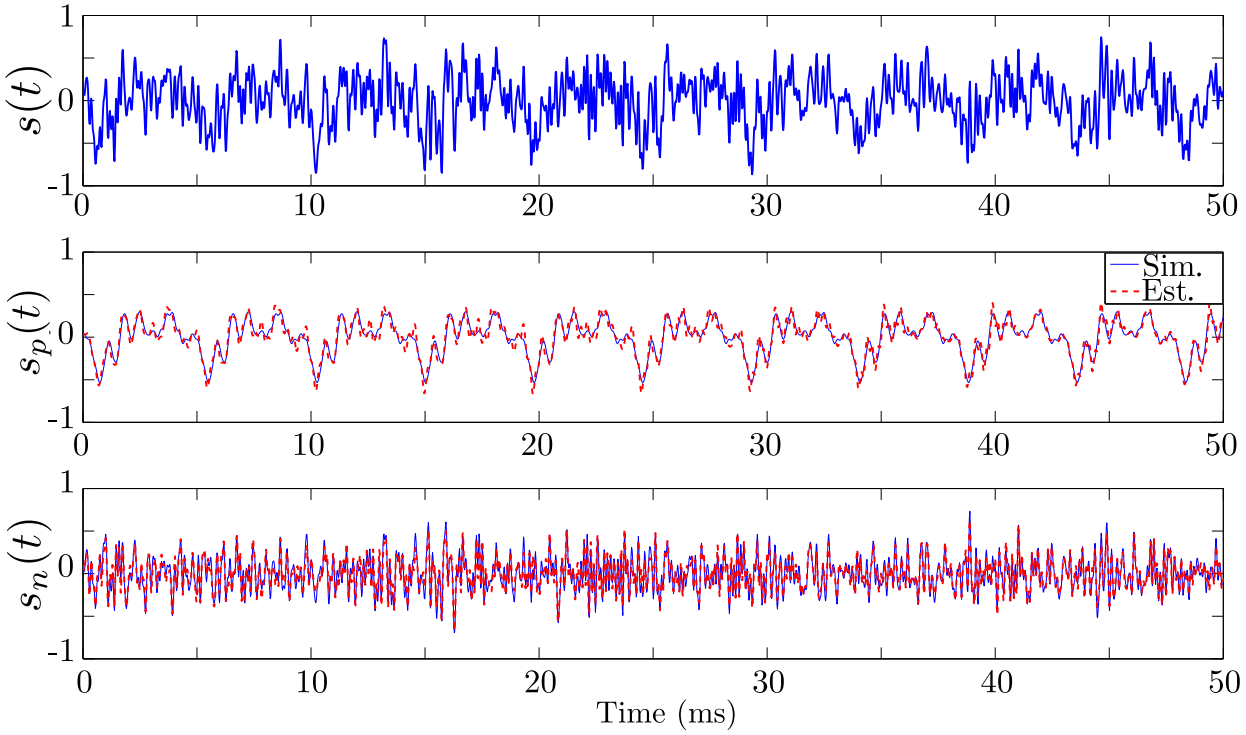

Figure 1: Results of the periodic/aperiodic separation of a simulated fricative /ʒ/. Voicing quotient is 50%, and $f_0 = 210$ Hz. Top is the mix signal $s(t)$, middle is the periodic signal $s_p(t)$, and bottom is the aperiodic signal $s_n(t)$. The estimated periodic and aperiodic signals are represented by a dashed line. Solid lines correspond to the simulated signals.

3.2 Results

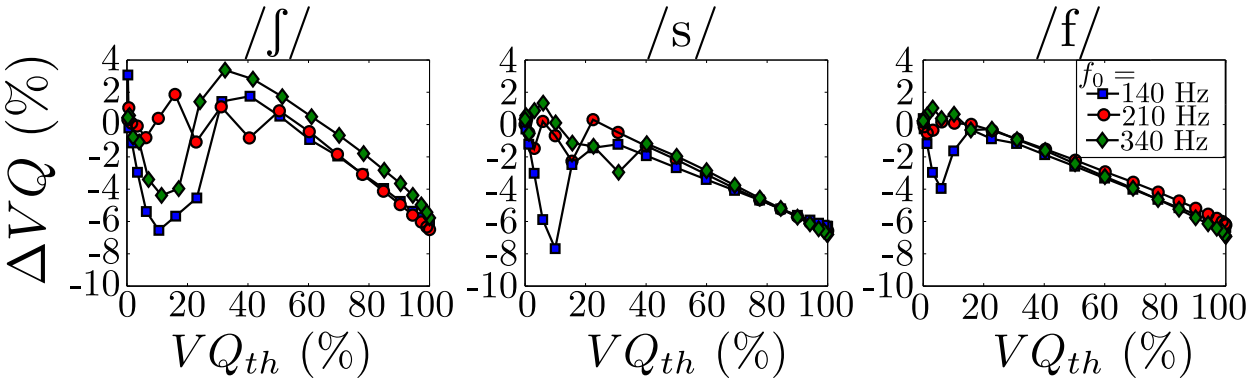

Figure 2: Error ΔVQ on the evaluation of VQ for $f_0 = \{140, 210, 340\}$ Hz as a function of the theoretical VQ, for the three simulated fricatives, /ʃ/, /s/, and /f/. Results are shown in %.

Fig. 2 shows the results of the simulations. $\Delta VQ = VQ_{est} - VQ_{th}$ quantifies the error of estimation on the voicing quotient VQ_{est} in regards with the theoretical voicing quotient VQ_{th}. When

the noise level is low, i.e. VQ is high, ΔVQ is negative and its absolute value increases as VQ increases. This means that when the periodic component is predominant, it is underestimated by the method. Possible explanation may be the fact that high-order partials are not detected and/or the energy of the detected partials is underestimated. In both cases, periodic components are still present in the separated aperiodic signal. Despite this limitation, the estimation error is very low and it can be applied to real signals with a confidence of 8% in worst cases, even in noisy conditions (low VQ). Note that, in the tested pitch interval, the method is robust to the number of considered periods.

4 Voice/noise separation of vowel-fricative-vowel sequences

4.1 Data and method

Analyzed data come from audio recordings of short pseudowords uttered by a French native male subject (54 y.o.). It consists in vowel-fricative-vowel sequences (VFV), where the vowel is /ɑ/, and fricatives are the voiced/voiceless pairs /ʃ,ʒ/, /s,z/, and /f,v/. Audio recordings have been acquired in a specific acoustically designed room, and at a sampling rate of 20000 Hz. For each of the 6 VFV sequences, the voice/noise separation algorithm presented in this paper has been applied, and the corresponding voicing quotient VQ has been computed.

4.2 Results

Results are shown in Fig. 3. It can be seen that the noise in the mix signals $s(t)$ has been successfully removed from the periodic signals $s_p(t)$, and that it corresponds to the separated noise signal $s_n(t)$. For instance, the voice signal during the voiceless fricatives vanishes, which is highlighted by a null voicing quotient. Interestingly, the evolution of VQ differs according to the uttered voiced fricative: for /ʒ/ and /z, VQ quickly drops down to a very small value, a few percents for /ʒ/, and around 20% for /z/, while VQ barely decreases for /v/, and constantly stays above 90%. This is then a useful tool to investigate the production of fricatives, for instance to quantify phonetic phenomena, such as voicing/devoicing assimilation.

5 Conclusions

This paper has presented a method for separating speech signals into a periodic component and an aperiodic component. In comparison with other methods, it is robust to high level of colored noise, thanks to an appropriate preprocessing applied to the raw speech signal. Numerical simulations have shown that the method successfully estimate the energy level of each component, even in the case of high level of noise. The error is always less than 8%, and is due to the underestimation of the periodic component that occurs when the latter is predominant over the aperiodic component. Applications on real speech signals have shown that the method is efficient to separate the voiced and frication noise components embedded in fricatives. The evolution of the voicing quotient, quantifying the amount of the voiced component in the speech signal, can then be computed. Results from a primary study showed that the

voicing quotient may be highly dependent on the uttered fricative.

Future works about the presented algorithm should include adaptive windows to ensure a sufficient number of considered periods, as proposed in [5, 12], especially in the case of very low pitched voices, as well as the use of automatic enumeration methods to estimate the number of harmonics embedded in the periodic signal. Since the periodic signal may be underestimated in case of low noise level, several iterations could be considered.

Acknowledgements

This study is partly supported by the ANR (*Agence Nationale de la Recherche*) ArtSpeech project.

References

[1] G. de Krom. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language, and Hearing Research*, 36(2):254–266, 1993.

[2] B. Elie and Y. 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*, 82:85–96, 2016.

[3] B. Elie and Y. Laprie. A glottal chink model for the synthesis of voiced fricatives. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5240–5244, March 2016.

[4] D. Erro, I. Sainz, E. Navas, and I. Hernaez. Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *Selected Topics in Signal Processing, IEEE Journal of*, 8(2):184–194, 2014.

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

[6] 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.*, 129(4):2144–2162, 2011.

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

[8] Y. Qi, B. Weinberg, N. Bi, and W. J. Hess. Minimizing the effect of period determination on the computation of amplitude perturbation in voice. *The Journal of the Acoustical Society of America*, 97(4):2525–2532, 1995.

[9] B. G. Quinn and P. J. Thomson. Estimating the frequency of a periodic function. *Biometrika*, 78(1):65–74, 1991.

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

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

[12] N. Sturmel. *Analyse de la qualité vocale appliquée à la parole expressive*. PhD thesis, Université Paris Sud-Paris XI, 2011.

[13] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *Speech and Audio Processing, IEEE Transactions on*, 6(1):1–11, 1998.

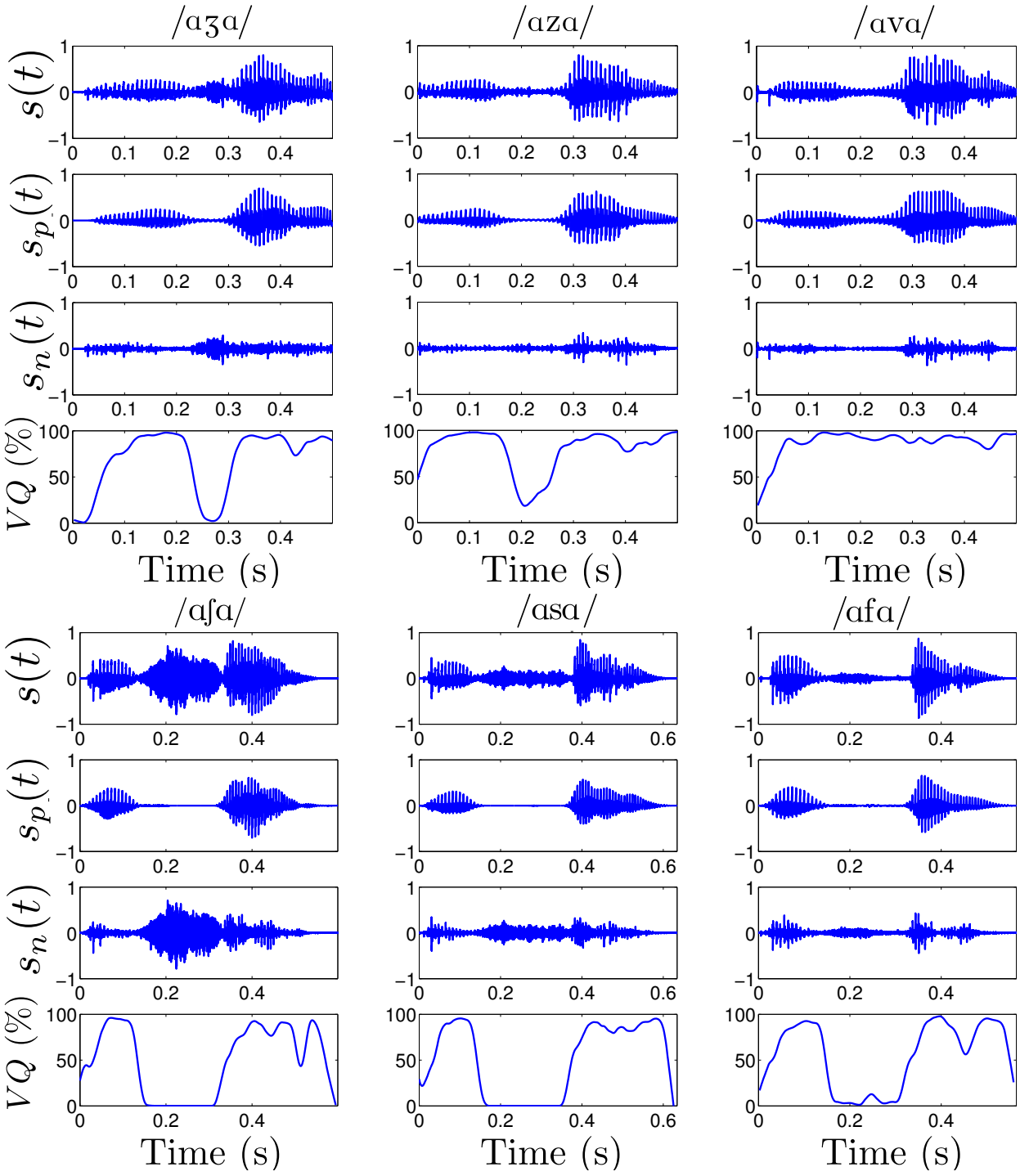

Figure 3: Results of the separation applied to VFV sequences. From top to bottom is the mix signal $s_(t)$, the periodic signal $s_p(t)$, the noise signal $s_n(t)$, and the voicing quotient VQ. From left to right is the palato-alveolar fricative /ʒ,ʃ/, the alveolar fricative /z,s/, and the labiodental fricative /v,f/.