USUAL TO PARTICULAR PHONATORY SITUATIONS STUDIED WITH HIGH-SPEED VIDEOENDOSCOPY

Gilles Degottex, Erkki Bianco, Xavier Rodet

IRCAM, Analyse-Synthèse
1, place Igor-Stravinsky, 75004 Paris
degottex@ircam.fr 0033 1 44 78 48 75

ABSTRACT

Current high-speed videoendoscopy (HSV) make it possible to obtain 4000 images of the larynx per second. By this process, the analysis of the vocal folds can provide significant information. This is also possible to estimate the area of the glottis. All this information is useful for the study of the various phonatory modes, but also for glottal flow estimation which allows the improvement of our acoustic understanding of speech signals. For the usual modes then for other particular phonatory situations, we present a comparison of various speech signals: acoustic, Electro-Glotto-Graphic, glottal area, and estimation of the glottal flow by inversion of the vocal tract.

Index Terms— high-speed camera, videoendoscopy, glottis, voice

1. INTRODUCTION

In the production of speech and singing, we are interested in the acoustic source signal produced by the glottis. The estimation of the glottal flow is a key problem for the modelling of the vocal production and its transformation. High-speed videoendoscopy is an increasingly common new technology. It allows for the estimation of the area of the glottis. With this estimation it is interesting to compare the glottal area with the various signals usually used in the study of voice production.

First of all, the various impacts of the physiology of the larynx on the acoustic source of the glottis can be observed on the videoendoscopic images. Among the observable phenomena, we can see important leakage (Fig. 6) and frequent lateral and longitudinal asymmetries [10]. In a such case, the closure instant of the vocal folds is difficult to define clearly, even on a HSV. The temporal synchronization of a model of glottal flow (LF[1], R++[2], etc.) with the acoustic signal is a recurrent problem in analysis of speech signals. This problem is usually simplified in a detection of the glottal closure instants (GCI). This instant is defined on the models as the instant of the strongest negative derivative. The flow at this instant may have a value bigger than zero. For all the phonatory situations, the following question arises: to what the GCI is associated with, in order to stay coherent with all the signals. From a detection of the GCI on the acoustic signal, we developed a method to estimate the shape of the glottal flow by inversion of the vocal tract [3, 4, 5, 6, 7].

Videoendoscopic images make it possible to follow the edges of the vocal folds [8, 9, 6, 10]. We have also developed a program to measure the glottal area [8]. The glottal area cannot be directly compared to the glottal flow. The glottal flow is not an immediate function of the area. Indeed, to obtain a flow from the area, it is necessary to solve a differential equation taking into account the impedance of the glottis and that of the vocal tract [11, 12]. This calculation is in progress.

Comparing the images and the area measure, the interpretation of Electro-Glotto-Graph (EGG) signal is improved.

2. MEASURES AND ESTIMATES

Glottal area estimate (in thick red line in the figures): The vocal folds are filmed through a rigid endoscope which passes through the mouth, connected to a high-speed camera *ENDOCAM 5562* which provides 4000 color images a second in 256x256 pixels. The glottal area is estimated by a thresholding method of the luminance of the videoendoscopic images [8]. The threshold is automatically computed by localising the edges of the vocal folds. The estimated glottal area is thus sensitive to the first visible lip of the vocal folds seen by the camera. This lip is not the same according to the opening and closing phase. During the opening phase of the glottis, the upper lip of the folds hides the lower lip. The estimated glottal area is thus focused on the upper lip. For example, in the figure 1, there is a delay between the maximum of the

EGG derivative and maximum of the glottal area derivative. This shows that a part of the vocal folds, certainly the lower lip, is moving while the glottal area does not change. On the contrary, during the closure, according to the phonatory mode and the importance of the Bernoulli's effect (see 3.1), the lower lip should close earlier than the upper lip. During this phase, the estimation of the glottal area can focus on the lower lip. Presently our method of area estimation does not allow us to obtain the constant of opening which expresses the leakage of the glottis. Over a period, the minimum of the estimated area is always zero.

Acoustic measure (thin and black lines): The acoustic signal is recorded with a sampling frequency of $44150Hz$ by a microphone placed on the endoscope, in front of the mouth, at $15.5cm$ from the head of the camera. During the recording, the head of the camera is in the back of the oropharynx. We estimate a distance oropharynx/larynx of $10cm$. The acoustic delay glottis/microphone is thus estimated to be $(0.155 + 0.10)/340 = 0.75 \cdot 10^3 s$. On figures, this delay is thus compensated (shown by a thin black interval on figures). We ask to the patient to pronounce /e/ for their comfort and to clear the field of the camera. The epiglottis is the most vertical with a /e/, except /i/ which is not possible to pronounce with endoscopy because of the position of tongue which is up against the palate. The pronounced phoneme sometimes changes to *schwa*. When a formant of the vocal tract get closer to the glottal formant, the phases and the amplitudes of this formant distort the glottal flow reproduced in the acoustic signal. The 1st formant of /e/ or /i/ is around $320 - 500Hz$. Knowing that the frequency of the glottal formant is between the 1st and 4th harmonic [5] (around $100 - 400Hz$), the influence of the 1st vocal tract formant on the glottal flow is not insignificant. It is thus very difficult to localize an instant of glottal closure directly on the acoustic signal.

Electro-Glotto-Graphic measure (EGG) (thick dotted blue lines): According to the recording, we used the device *EG90* of *F-J Electronics* for figures 1 in 8 and the device *Laryngograph* (former version) of the company *Laryngograph* for figures 9 and 10. The signal is sampled in perfect synchronization with the acoustic signal with a sampling frequency of $44150Hz$. The EGG is finally high-pass filtered with a cut of frequency of $40Hz$ to remove the influences of the larynx movements. The synchronization delay between the images and the EGG is at most 3 images what implies a maximum delay of $0.75ms$ (shown by the same thin black interval as the glottis/microphone delay).

Glottal flow estimate (thick dotted green lines (grey in black&white printing)): By following the source-filter hypothesis, the parameters of a LF-model [1] is estimated from the acoustic signal by a method of vocal tract inversion [3]. At first, the glottal closure instants are estimated then the shape of the wave of the glottal flow. The model of the vocal tract is an all-pole stable filter, and thus of minimal-phase. It does not model the delay of propagation in the vocal tract. On the figures, this delay is also compensated for as on the acoustic signal. The effective glottal closure instant as the instant of the strongest impulse of the glottal source can be badly defined. For example because of the leakage of the glottis or the asymmetry of the movement of vocal folds. Because our temporal synchronization of the glottal flow model is based on this unique instant, the whole model can be badly placed (Fig. 1 4 5 6). We work at present on a joint estimate of the parameters of the model. Computing so in this way, the synchronization and the shape in the same time.

3. PRESENTED SITUATIONS

For each figure: the left column shows a videoendoscopic image of the glottis with the posterior part at the top. In the right column, the top graph shows the area in the thick red line, the EGG in the thick dotted blue line and the estimation of glottal flow in the thick dotted green line (except for the case of the exhaled fry). The bottom graph shows the acoustic signal in the fine black line and the derivatives of the previous mesures and estimates. For a better readability, amplitudes of the signals are normalized to their standard deviation.

The most usual phonatory modes

3.1. Mode *I* (Fig. 1, 2 and 3)

The mode *I* is the most common phonatory mode used in speech. The minimum of derivative of the EGG signal seems to well coincide with the minimum of the derivative of the area (*i.e.* the moment when vocal folds move the fastest). On the other hand, this coincidence does not appear during the opening of the glottis. The EGG thus reveals a behavior hidden by the upper lips of the vocal folds.

According to literature, vocal folds separate gradually and close quickly. The Bernoulli's effect explains this effect at the closure by an aspiration of the upper lips of the vocal folds. However, this effect is not systematic, the figure 1 shows the case we mostly observed: an opening nearly as fast as the closure.

The figure 2 illustrates Bernoulli's effect. In that case, we can notice that the opening is made of two steps. Maybe because the vocal folds are made of layers. This causes *ripples* on the glottal area. During the closure, 4000 images per second limit the precision of estimation of the movement of vocal folds.

Currently, the EGG of the figure 2 seems too much difficult to interpret. It has thus been removed.

3.2. Mode *II* (Fig. 4 and 5)

The mode *II* is usually present in a high-pitched voice. In that case, we notice generally that vocal folds are more tightened and more parallel. Vocal folds are in movement throughout the period. Therefore, the duration of the closed phase in a model of glottal flow should be zero. The glottal area is more like a sine curve and the derivative of the area supports this proposition. In the opposite way of the mode *I*, only the vocal fold cover is moving and so, there is only one lip. We can thus consider that the glottal area estimate well describe the effective glottal area.

Particular phonatory situations

3.3. *Breathy* voice (Fig. 6)

Acoustically, the breathy voice is defined by the presence of air in the sound. But there is not necessarily creaky noise. The videoendoscopic image of the figure 6 shows, over a period, the most closed glottis. The leakage on the posterior part of the vocal folds is evident and omnipresent in this situation. This leakage generates the turbulence noise of this voice. The posterior part is practically immobile while the anterior part generates a harmonic sound. A part of the glottis thus remains opened throughout the period. The resonances are less marked because the bandwiths of the formants are bigger in the case of an opened glottis which provokes a supplementary acoustic loss. As in mode *II*, vocal folds are also always in movement. This implies a closed phase of the glottal flow to be zero and a shape of the glottal area getting closer to a sine curve.

3.4. *Tense* voice (Fig. 7)

Physiologically, it is an excess of pressure of the ventricular folds, without reaching the critical case of pressed voice. Along the *breathy*, *tense* and *pressed* voices, we can define a scale of relaxation-tension. Acoustically, this scale is defined by a ratio between the level of noise and the level of the harmonic sound. On this scale, the *tense* voice is opposite to the *breathy* voice. The closed phase of the *tense* phonation is longer than in *breathy* phonation. Compared to the *breathy* voice, the EGG shows that the distance between the maximum and the minimum of his derivative is correlated to the opened phase duration, without being equal to it. The minimum of the EGG derivative and the minimum of the area derivative seems to coincide. This is not the case for the maxima.

3.5. *Pressed* voice (Fig. 8)

The false vocal folds and the interarytenoid muscle interfere with the vocal folds when opening and restrain them from making an ample movement. The vocal folds remain somehow stuck to one another while allowing brief air impulses to pass trough. In these conditions, the computation of the glottal area does not work. This voice is produced by a tightening of the ventricular folds. By reflex action, a space between the vocal folds can exist. On the scale of relaxation-tension (see 3.4) this voice is a degenerate case of extreme tension. The EGG shows a movement of the masses of the larynx while the glottis is not visible.

3.6. Exhaled and inhaled *fry* (Fig. 9 et 10)

The *fry* voice (often called mode 0) is defined by creaky quality, a low fundamental frequency ($\sim 50Hz$), a short opening phase, a fast closure and irregular instants of closure. Because of this irregularity, this voice is nearly impossible to study by stroboscopy. By comparing the EGG and the glottal area (Fig. 9 and 10 at the top right), we can notice that the EGG does not reveal movement of the vocal folds which is not shown by the area. The variations of the EGG are temporarily synchronized with those of the area. The visible lips seen by the camera defines relatively well the whole movement of the vocal folds. Moreover, the mass of the vocal muscle does not move, indicating that only the upper lips move and close the glottis. The exhaled *fry* is made by a relaxation of the larynx and a subglottal pressure lower than in a usual phonation. The inhaled version is made by a tension of the vocal folds and a contraction of the diaphragm which aims to create a depression in the lungs. The impulses made by the opening of the glottis can be regular enough to perceive a pitch. The reached intensity is more important thanks to the short opening phase and the very important depression. Contrarily to the exhaled version and surprisingly, only the lower lips close the glottis.

4. CONCLUSIONS

Physiologically, we can notice that the posterior part of the glottis can remains opened while the anterior generates a harmonic sound (Fig. 6). Furthermore, depending of the significance of the Bernoulli's effect, the videoendoscopic images and the estimation of the glottal area show us that the vocal folds do not close necessarily faster than they open (Fig. 4, 5 and 6).

Concerning the glottal area: because of the coupling between the vocal tract and the glottal source, *ripples* appears on the glottal flow. We can see that *ripples* can also appears on the glottal area (Fig. 2). Moreover, the glottis does not stop varying in mode *II* and breathy phonation. The closed phase duration of the glottal flow should thus be zero.

By comparing the EGG and the glottal area, we can notice the following elements: the instant of strongest negative derivative of the EGG seems to correspond with the strongest negative derivative of the area. This is the instant when the vocal folds close the fastest but not necessarily when they touch each other. Moreover, The maximum of the EGG derivative reveal behaviors of the masses which are hidden by the visible lips seen by the camera. The measured duration between the maximum and minimum of the EGG derivative seem proportional to the maximum-to-minimum duration of the glottal area derivative (duration between the fastest opening instant and the fastest closure instant). However, this duration is not equal to the effective contact duration of the vocal folds.

The measure of the acoustic pressure, after synchronization with the other signals, shows us that the strongest acoustic depression does not correspond necessarily to the instant of effective closure of the vocal folds. Furthermore, the estimation of the glottal flow does not seem synchronized with the other signals as long as its shape and its temporal synchronization are not jointly estimated.

5. THANKS

We make a point of thanking the *Richard Wolf* company for lending us their full system of high-speed camera *ENDOCAM 5562*, *F-J Electronics* for lending us an *EG90* Electro-Glotto-Graph, Laurier Fagnan for his contribution to the *inhaled fry* and Mette Pedersen for her assistance in the recordings.

6. REFERENCES

[1] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[2] R. Veldhuis, "A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation," *JASA*, 1998.

[3] G. Degottex and X. Rodet, "Voice source and vocal tract separation," *to be published*, 2008.

[4] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of lf glottal source parameters based on an arx model," *Interspeech*, 2005.

[5] Nathalie Henrich, *Etude de la source glottique en voix parlee et chantee*, Ph.D. thesis, UPMC, 2001.

[6] H. Pulakka, "Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography," M.S. thesis, Helsinki University of Technology, 2005.

[7] R. Fernandez, *A Computational Model for the Automatic Recognition of Affect in Speech*, Ph.D. thesis, Massachusetts Institute of Technology, 2004.

[8] G. Degottex, E. Bianco, and X. Rodet, "Measure of glottal area on high-speed videoendoscopy," *to be published*, 2008.

[9] Deliyski and Petrushev, "Methods for objective assessment of high-speed videoendoscopy," *AQL*, 2003.

[10] J. Neubauera, P. Mergell, U. Eysholdt, and H. Herzel, "Spatio-temporal analysis of irregular vocal fold oscillations: Biphonation due to desynchronization of spatial modes," *JASA*, 2001.

[11] J.L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer Verlag, 1972.

[12] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, 1982.

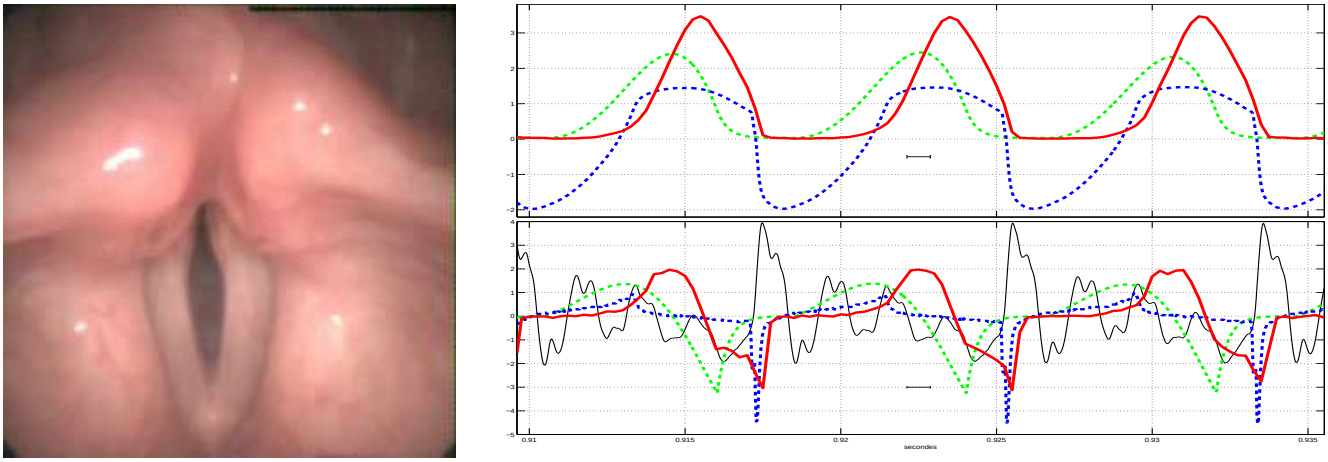

Fig. 1. Mode I of a man

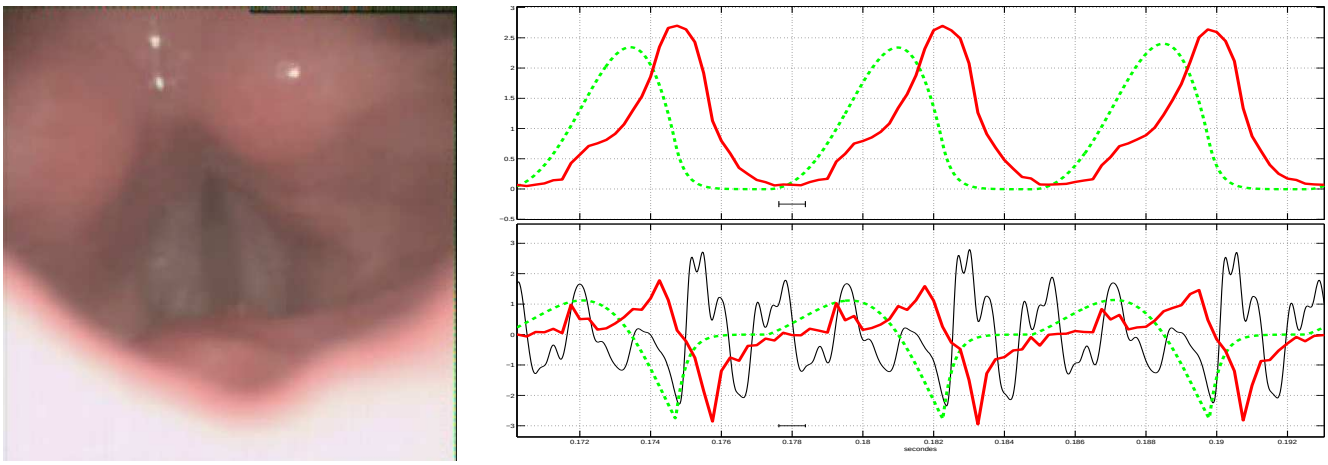

Fig. 2. Mode I of a man with fast closure

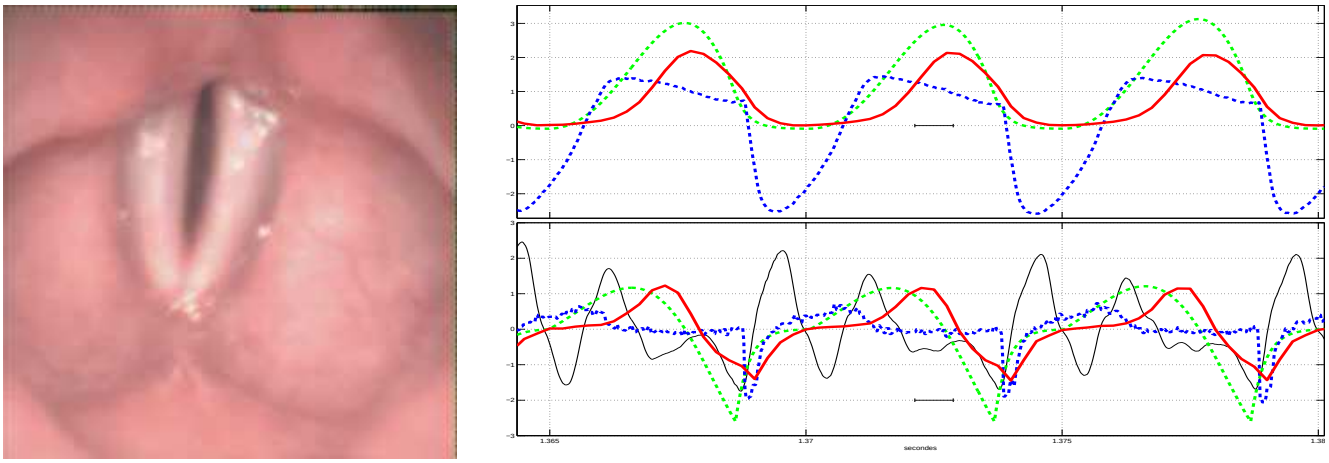

Fig. 3. Mode I of a woman

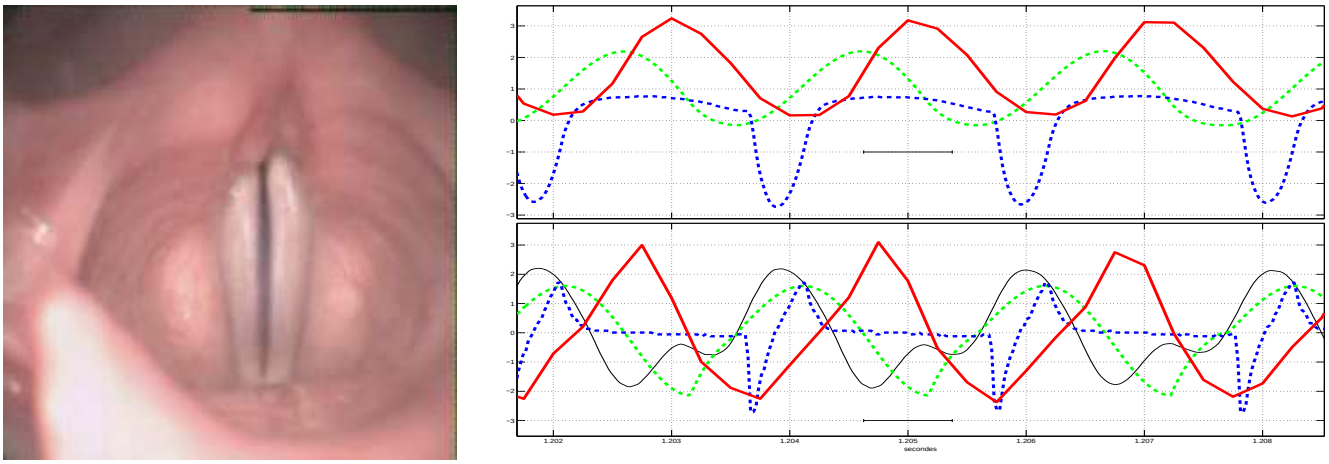

Fig. 4. Mode II of a woman

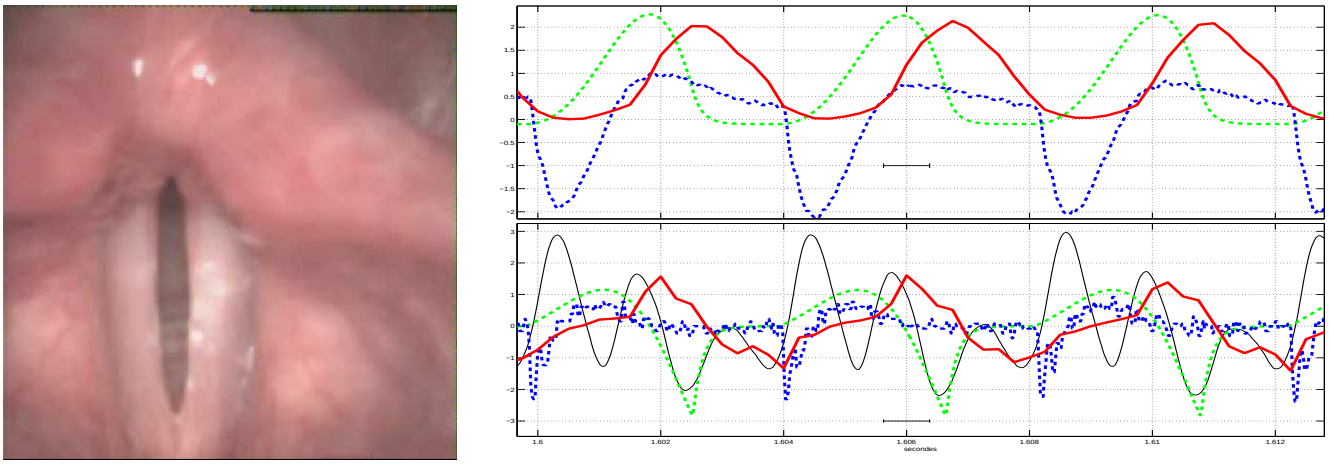

Fig. 5. mode II of a man

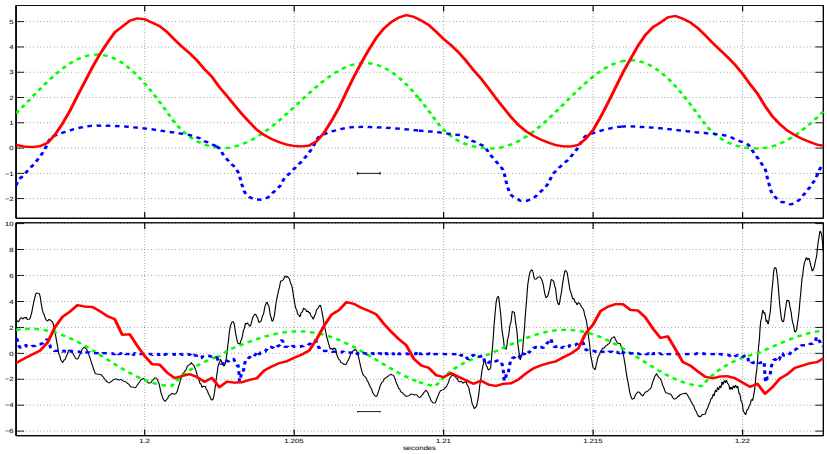

Fig. 6. Breathy voice

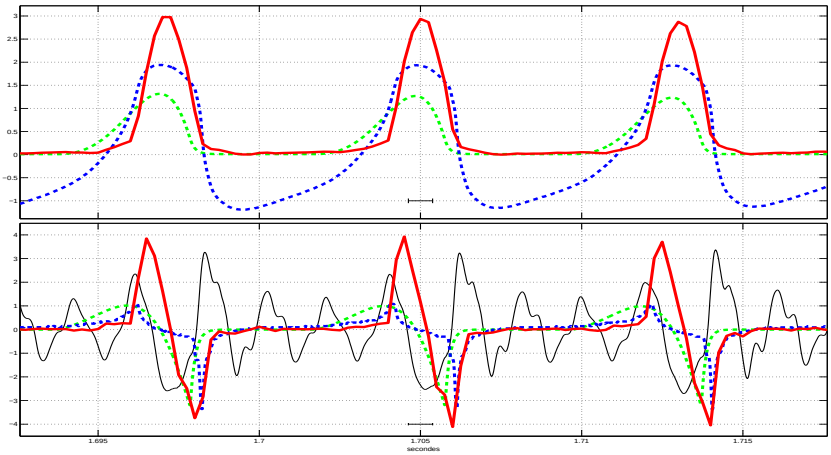

Fig. 7. Tense voice

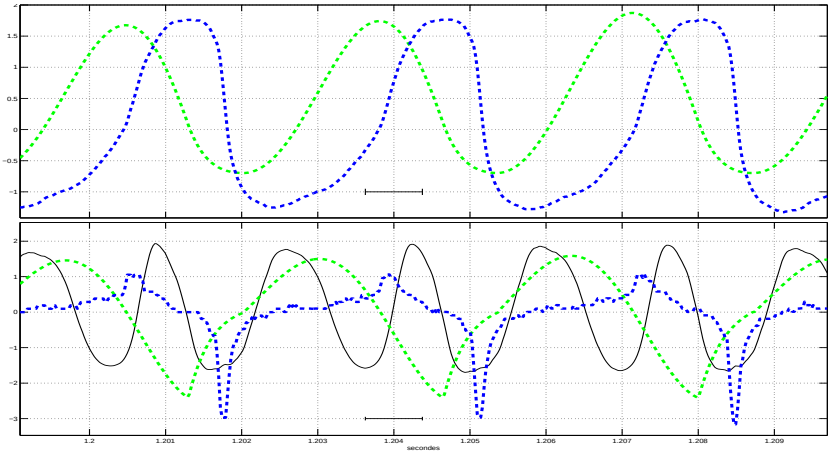

Fig. 8. Pressed voice

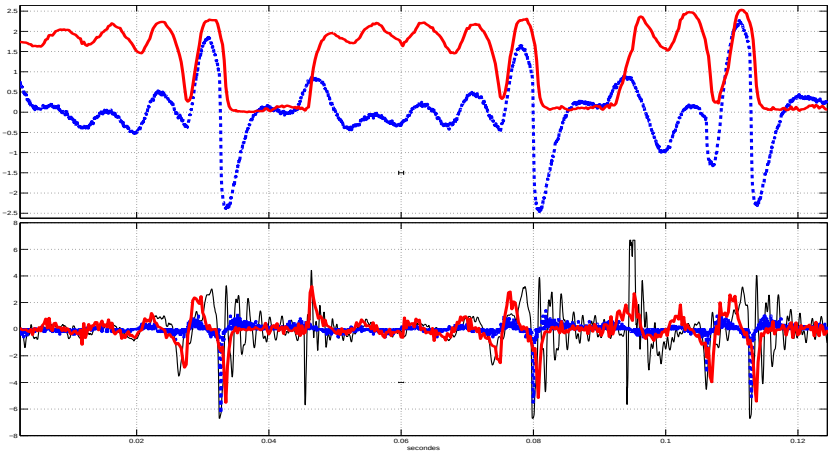

Fig. 9. Exhaled Fry

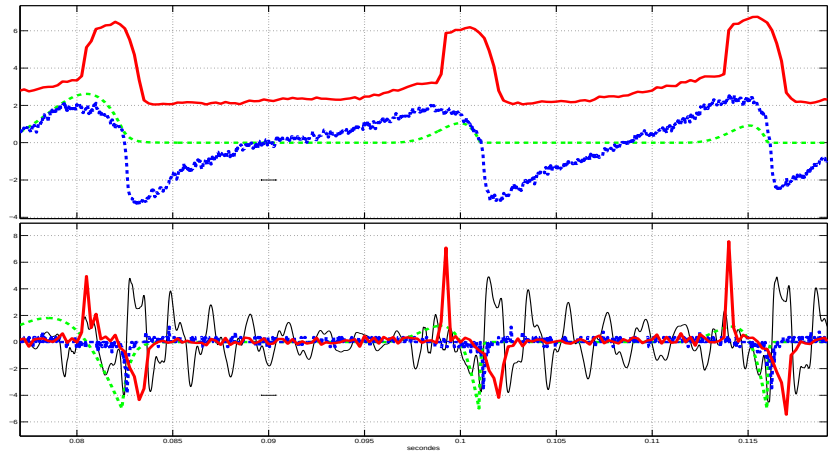

Fig. 10. Inhaled Fry