

## Vocal tract acoustic transfer function measurements: further developments and applications

Y. PHAM THI NGOC and P. BADIN

*Institut de la Communication Parlée, URA 368 du CNRS, INPG, Université Stendhal, 46 Avenue Félix Viallet, 38031 Grenoble cedex 01, France*

**Abstract:** Recent techniques of evaluation of vocal tract acoustic transfer functions are based on the external excitation of the tract at the thyroid cartilage level. This paper presents further developments of methods using Gaussian white noise or pseudo-random sequences. In addition to the traditional non-parametric spectral characterisation, an automatic estimation of formants and bandwidths has been developed using model identification procedures. A user-friendly protocol has been implemented to perform quasi simultaneous recording of transfer functions and radiated sound. This method has been used to study the influence of glottis conditions on formants and bandwidths for the French vowels, and an experimental *formant-cavity affiliation index* has been defined as the relative bandwidth variation between closed glottis and sound production conditions. Finally a first attempt to characterise the aspiration noise spectrum for whispered vowels has shown that the overall source spectral tilt increase with overall SPL.

### 1. INTRODUCTION

A very interesting way to characterise vocal tract acoustics is to measure acoustic transfer functions directly on subjects. One of the important advantages of this characterisation in the frequency domain is the possibility, when operating in closed glottis conditions, to measure the acoustic transfer function without influence of the sources. More generally, the possibility to measure the transfer function without resorting to the speech signal itself ensures the quality of the measurements, and allows to study more precisely the basic geometric/acoustic relationships. This paper presents the last developments of such acoustic transfer function methods, and a few examples of applications to speech.

### 2. VOCAL TRACT ACOUSTIC TRANSFER FUNCTION MEASUREMENT

#### 2.1 General principles – Parametric and non-parametric characterisation

The acoustic transfer function of the vocal tract is defined as the complex ratio between the acoustic flow excitation at the glottis and the acoustic flow at the lips or nostrils. The basic measurement technique consists in exciting externally the vocal tract with the help of a small shaker at the thyroid cartilage level, and picking up the resulting signal with a microphone. The measurement chain (cf. [1] for more details) consists of a signal processing board that generates the excitation signal feeding a mini-shaker (BK4810) through a D/A converter and power amplifier, and of a microphone connected to the A/D channel of the board. The total transfer function  $T_{tot}(f)$  between the D/A and A/D channels of the board can be written  $T_{tot}(f) = E(f) \cdot N(f) \cdot T(f) \cdot R(f)$ , where  $E(f)$  is the frequency response of the excitator,  $N(f)$  the transfer function of skin and cartilages of the neck,  $T(f)$  that of the vocal tract, and  $R(f)$  the radiation characteristics at the lips. As the major goal is to determine  $T(f)$ , knowledge of the other components is needed.

A method based on a Gaussian white noise excitation (henceforth WNE) has been developed by [2]. More recently, the method based on a pseudo-random excitation (henceforth PRE), employed in room acoustics, has been applied to the vocal tract by [1]. With these methods, either parametric or non-parametric model estimations can be performed. Non-parametric estimations result in a description of the transfer function as a vector of amplitude and phase against frequency, whereas parametric estimations characterise the transfer function in terms of complex poles and zeros.

The non-parametric estimation of the transfer function with the WNE method (cf. [2]) simply consist in

evaluating the long time average spectrum of the output signal: it requires several seconds of sustained production of the vocal tract configuration under investigation. In the PRE method, the property of the pseudo-random sequence of having an autocorrelation identical to a single impulse, and the assumption that this pseudo-random sequence is not correlated with the voice or noise sources generated during actual sound production, lead to compute the transfer function as the FFT of the crosscorrelation between the pseudo-random sequence itself and the output signal (cf. [1]). This method allows short measurement durations and actual sound production during the measurement.

The theory of model identification presented in [3] allows a parametric estimation of the vocal tract transfer function. If the system to be characterised is linear, the relation between an input signal  $u(t)$  ( $t=1, 2, \dots, N$ ), an output signal  $y(t)$ , and an additional, unmeasurable disturbance (noise)  $v(t)$  is:

$$y(t) = \sum_{k=1}^{\infty} g(k) \cdot u(t-k) + v(t), \quad \text{with} \quad G(q) = \sum_{k=1}^{\infty} g(k)q^{-k}, \quad \text{and } q \text{ being the delay operator.}$$

$G(q)$  is the system transfer function, and  $g(k)$  the corresponding impulse response. The disturbance  $v(t)$  can be described as a white noise  $e(t)$  filtered with the function  $H(q)$ . In the case of an ARX model,

$$G(q) = q^{-nk} \frac{B(q)}{A(q)} \quad \text{and} \quad H(q) = \frac{1}{A(q)}$$

where  $B$  and  $A$  are polynomials of respectively  $n_a$  and  $n_b$  order in  $q^{-1}$ , and  $nk$  is the number of delays from input to output. With the help of the software package MATLAB, ARX identifications have been performed for the two types of excitation signals. The first step, i.e. the choice of optimal model order, is achieved using Akaike's Information Theoretic Criterion [3]. Then the ARX model parameters, i.e. the coefficients of the polynomials  $A$  and  $B$ , are estimated by a least squares method. The corresponding transfer function is displayed, and the formants and antiformants are computed from pairs of complex conjugate roots of the polynomials defined by  $z_k, z_k^* = e^{-\sigma_k T} \cdot e^{\pm j 2\pi F_k T}$ , where  $T$  is the sampling period,  $F_k$  is the pole or zero frequency, and  $B_k = \sigma_k / \pi$  is the bandwidth.

## 2.2 Evaluation of the measurement chain

In this section, the frequency characteristics of the different components of the measurement chain are presented and discussed.

The mini-shaker is constituted of a mobile element suspended by a spring to a massive cage. It is designed in such a way that the first resonance of the free system is very low ( $\approx 53$  Hz). In order to ensure a good contact between the shaker and the neck skin, a small aluminium disk is screwed to the shaker table (cf. [1]): it has been verified theoretically that the effective dynamic mass of the mobile element and spring stiffness of the table loaded with the disk and the neck skin lead to a resonance frequency of about 40 Hz. This frequency is thus very low and depends little on the load of the shaker. Another resonance around 18 kHz, corresponding to the first axial resonance of the shaker mobile element, is high enough not to perturb the 0-10 kHz frequency range of interest for speech. Using a small accelerometer and the WNE method with a sampling rate of 40 kHz, it has been experimentally verified that the frequency responses of the system with and without skin load are almost identical, within  $\pm 5$  dB, up to 10 kHz. Thus, the frequency response of the shaker loaded with the neck skin of a subject (see Fig. 3) is used as a correction curve.

It is extremely difficult to realise an acoustic flow source having an internal impedance high enough to ensure the independence of its output signal in relation with the acoustic impedance loading the source. It is even more challenging to envisage to insert such a source in a subject vocal tract near the glottis. This is why the vocal tract is externally excited. As a result, the unknown transfer function  $N(f)$  of the skin, tissues and cartilage in this region of the neck is added to the measured transfer function. In order to measure the frequency transmission characteristics of the neck tissues, one should be able to determine acoustic flows inside the vocal tract. A direct measure is impossible in practice, as only pressures could be measured inside the tract with miniature microphones. The complex pressure in the vicinity of the glottis is the product of the flow at the same location by the input impedance of the tract seen from this point, looking towards the lips. As this input impedance is unknown (its poles are the same as those of the transfer function), an indirect measurement of the acoustic flow at the level is impossible. We have verified experimentally, both on a plexiglass tube and on a subject (using a nasally inserted microphone probe), that the pressure signals measured inside the tract are coherent with this hypothesis. Assumptions must thus be made about  $N(f)$ : it is very likely that this function has not sharp poles nor zeros, and has a global low-pass filtering effect only.

Due to the low level of output signal, the microphone must be positioned close to the lips, and thus in the near field of the radiating region, where the relation between radiated pressure and acoustic flow is not well determined. This problem can however be overcome for the determination of the excitation source spectra: since the speech sound and the signal resulting of the external excitation of the tract are picked up with the same microphone at the same location, the influence of the radiation characteristics  $R(f)$  cancels out.

### 2.3 Comparison of the WNE and PRE methods

Both methods have already been used and shown to lead to similar results for human subjects [1]. However, no formal experimental comparison has been made so far. Thus, plexiglass tubes, of sizes comparable to that of a human vocal tract, have been characterised with both methods, in identical experimental setup conditions. The output side of the tubes were mounted flush in a plexiglass baffle. The input side were fitted with a rubber membrane which was hooked to the shaker table by means of a threaded rod screwed into the table and of two nuts pinching the membrane, one on each side of the membrane pierced in its centre by the rod. The sampling frequency was set to 10 kHz. A number of 2048 points of input and output signals were used for both methods. An optimal number of 12 poles and 12 zeros was found for the WNE method, and one more pole was found with the PRE method. The differences between the resonance frequencies never exceeded 0.5%, and the differences between the bandwidths were always lower than 10%. This shows that the methods are equivalent, with, however, an advantage to the PRE method that can yield a non-parametric characterisation even in sound production conditions.

## 3. APPLICATIONS

### 3.1 Experimental protocol

In order to ease different types of experiments and allow the acquisition of sizable transfer function corpuses for different subjects, a modular user-friendly protocol has been developed. An example of basic protocol is depicted in Fig. 1 for the case of a quasi-simultaneous acquisition of a transfer function in sound production condition and of the sound produced with the same articulation. The arrows at the bottom of the figure mark the keyboard strokes that allow the subject to start the different phases. During phase 1, the subject positions the excitator by listening to his/her own articulation heard like whispered speech. Sound production is started during phase 2. Phase 3 is the measurement with the PRE signal, and the sound itself is recorded during phase 4. The rest of this section will exemplify different combinations of basic protocols. Finally, it should be mentioned that a computer programme prompts the subject with the different instructions and corpus items in an automatised manner.

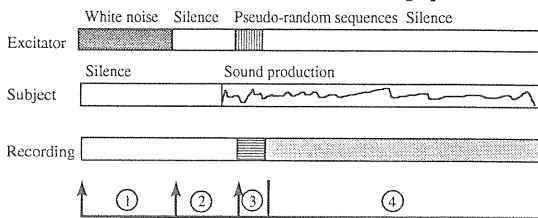


Fig. 1 – Example of recording protocol

### 3.2 Influence of glottis state on formants and bandwidths for vowels

The conditions at the glottis are known to affect the formants and bandwidths of the vocal tract acoustic transfer function. In order to evaluate this influence, the transfer functions of the eleven French oral vowels have been recorded by five subjects (four males, one female). For each of the twelve repetitions recorded for each vowel, the subjects were instructed to take the articulation for the vowel, then to close their glottis for the first measurement, and finally – without moving the other articulators – to start voiced sound production to perform the second measurement. With these experimental conditions, it seems reasonable to assume that the only difference between the two measurements is related to the difference in glottis condition only. After removal of the faulty measurements, the average per subject et per vowel of the formant frequencies and bandwidths in both glottis condition has been computed. As well, the relative variations of the frequencies and bandwidths have been systematically evaluated.

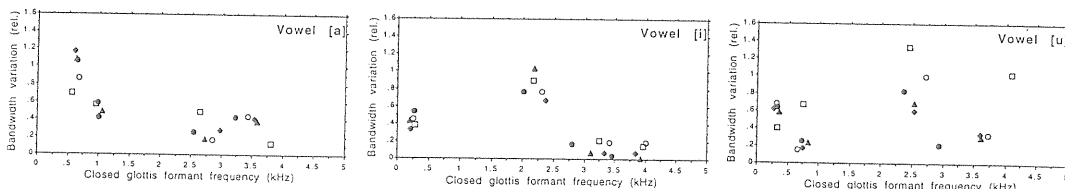


Fig. 2 – Bandwidths variations against closed glottis formant frequencies (HL□, JLS○, MP◆, PB▲, CS○)

It appears that the formant frequencies are not very much affected by the glottis state, but for F1. For [i] and [y], F1 systematically increases by more than 10% in sound production condition. This large variations could be ascribed, in addition to the glottis influence itself, to the fact that the subjects may have maintained too small constriction sizes during the closed glottis measurements (due to the difficulty to control very precisely the tongue tip position, in absence of the kinaesthetic feedback that is usually provided by the air flow through the vocal tract), and thus too low vocal tract first resonances (in the case of these vowels, the first vocal tract resonance is the Helmholtz resonance between the constriction and the cavity behind it).

The situation is very different for the bandwidths. Results for [a, i, u] are depicted in Fig. 2. The notion of “bandwidth retention”, related to formant-cavity affiliations, has been theoretically discussed by [4]. The present experiment provides a rather good mean to determine if a given formant is affiliated or not to the back cavity: if the bandwidth associated with this formant differs significantly in closed glottis and sound

production conditions, then the formant is principally affiliated to the back cavity; if the bandwidths are rather similar, then the formant is principally affiliated to the front cavity. The percentage of increase of the bandwidth from the closed glottis to the sound production condition appears to be an interesting experimental *affiliation index*. It is out of the scope of this paper to discuss the results in more detail, but it worth mentioning the clear cut answer to the question of affiliation of F1 and F2 of [u]: for the male subjects, F1 is the back cavity / oral constriction Helmholtz resonance, and F2 the front cavity / lip constriction Helmholtz resonance; for the female subject (HL), it is the opposite. The knowledge of these affiliations will help better understanding the articulatory-acoustic relationship, in particular for inversion purposes (cf. [5]).

**3.3 Aspiration source spectra for whispered vowels**

In order to determine the spectrum of excitation sources for whispered vowels, we have carried out simultaneous sound and transfer function recordings on the French whispered vowels [a, ε, œ] sustained by one subject, at different effort levels. Since the aspiration source is located near the glottis, no zero appears in the transconductance between this series pressure source and the acoustic flow at the lips. Thus, no zero should be found in the sound spectra themselves, which was verified in the data. The source spectra can thus be obtained by simply subtracting the measured transfer function from the sound spectra. The correction for the shaker characteristics has been taken into account, but no correction has been used for the neck transfer function: all the source spectrum data obtained include the gentle low-pass filter effect supposed for this unknown neck characteristics. An example of associated sound spectrum, transfer function and aspiration spectrum is given in Fig. 3. The aspiration spectrum appears to be relatively flat in the region 50 Hz – 4 kHz; an average spectrum tilt has thus been estimated in this range by linear regression. Fig. 4 shows that, for the three vowels, this tilt increases with increasing effort level, i.e. increasing overall SPL. An increase of the overall spectral tilt for increasing SPLs had already been noted by [6] for voiceless fricatives.

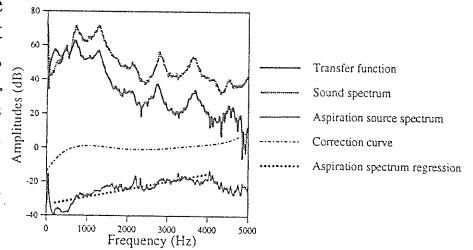


Fig. 3 – Spectra for vowel [a] of PB

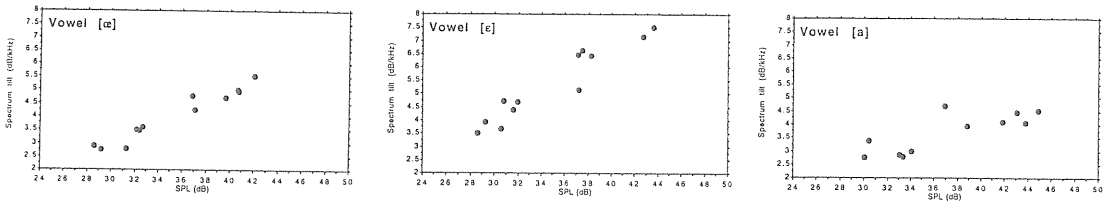


Fig. 4 – Aspiration source spectra tilt against overall SPL levels (subject PB)

**4. CONCLUSIONS AND PERSPECTIVES**

Methods for measuring vocal tract acoustic transfer functions have been presented. A new feature is the possibility to detect automatically formants and bandwidths, using model identification. The potentialities and limits of these methods have been shown and discussed. Different corpuses have been recorded. The influence of the glottis condition on formants and bandwidths has been shown, and an *affiliation index* has been defined and applied to the French oral vowels. A first investigation of aspiration and spectra for vowels has shown that the aspiration source spectrum tilt increases with increased effort level. In the near future, data on vowels and fricatives will be recorded, in order to improve source models and to develop better articulatory-acoustic models.

**Acknowledgements**

This work has been partly supported by the French CNRS and the European ESPRIT project *SPEECH MAPS*.

**References**

- [1]Djeradi A., Guérin B., Badin P., and Perrier P. (1991), "Measurement of the acoustic transfer function of the vocal tract: a fast and accurate method", *J. of Phonetics* 19, 387-395.
- [2]Castelli E. and Badin P. (1988), "Vocal tract transfer functions measurements with white noise excitation. Application to the naso-pharyngeal tract." 7<sup>th</sup> FASE Symposium, Speech 88, Edinburgh, 415-422.
- [3]Ljung L. (1987), *System identification – Theory for the User*, Prentice-Hall: Englewood Cliffs, N.J.
- [4]Badin P., Perrier P., Boë L.J., and Abry C. (1990), "Vocalic nomograms: acoustic and articulatory considerations upon formant convergences", *J. Acoust. Soc. Am.*, 87(3), 1290-1300.
- [5]Baillly G. (1993). Resonances as possible representation of speech in the auditory-to-articulatory transform. In *Proceedings of Eurospeech 93*, Berlin, 1511-1514.
- [6]Badin P. (1989). Acoustics of voiceless fricatives : production theory and data. *STL-QPSR* 3/1989, 33-55.