Emotional Voice Measurement : A Comparison of Articulatory-EGG and Acoustic-Amplitude Parameters

Solange Rossato, Nicolas Audibert & Véronique Aubergé

Institut de la Communication Parlée
Université Stendhal/INPG/CNRS, Grenoble, France
{rossato;audibert;auberge}@icp.inpg.fr

Abstract

NAQ has been proposed as the 4th prosodic dimension of expressive speech [5]. This paper aims at testing the consistency, for characterizing emotional expressions in voice, of the Normalized Amplitude Quotient (NAQ, [1]) vs. the estimated Open Quotient (OQ, [10]) parameter vs. the direct EGG measurement of glottal parameters. Those parameters were tested on an authentic expressive speech corpus [3]. The phonemic influence of the NAQ parameter was first evaluated by matching measure locations with an expert phonetic labeling. Estimations of F0 and OQ calculated on the one hand by inverse filtering and on the other hand from electroglottography (EGG), were then systematically compared. Results show a speaker-dependent phoneme effect on NAQ, and seem moreover to indicate a systematic overestimation of NAQ on [n] segments. In parallel, the comparison between inverse filtering and EGG parameters shows an underestimation of F0 used for the calculation of amplitude-based parameters. No correlation could be found between the OQ values calculated from both methods.

1. Introduction

Voice quality, both from objective and impressionistic criteria, has been related to the vocal expression of affects, together with other extra-linguistic information such as the speaker's age and sex. Some psychological studies have lead to the integration of vocal expressions, including voice quality features, in a comprehensive model of the production of emotion mainly based on acted emotional speech. According to Scherer et al.'s proposals [13], the general tension state of the larynx muscles is assumed to be affected directly by the emotional response, which implies that voice quality participates in the expression of emotion. Additionally, variations of spectral parameters related to voice quality such as spectral slope are predicted by the model, together with other prosodic parameters. For instance, spectral slope is predicted to increase for sadness, and to decrease for anger.

Perception experiments of synthesized stimuli [8], varying on voice quality only, showed the effect of glottal source variation on attitudinal and emotional speech.

Laver [11] proposes a comprehensive description of laryngeal muscular settings associated with resulting voice qualities impressionistically labeled and suggests, for English, breathy voice to be linked with intimacy, whispery voice with confidentiality and harsh voice with anger. Campbell [5] points out the correlation between the degree of "care" in the voice, and the pressed-breathy continuum, that he describes as the variation of the logarithm of Normalized Amplitude Quotient (NAQ) proposed by Alku [1], independently of F0 variations.

This paper aims at testing the consistency of the calculation of NAQ parameter for characterizing emotional expressions: an NAQ algorithm, developed by Mokhtari [12], has been applied to a phonetically balanced corpus, on two different speakers, for different authentic emotional expressions, in order to verify the phonemic robustness of this voice quality parameter. The NAQ is an estimation of the duration of the glottal closing phase. In order to get the closest reference to NAQ, the Open Quotient (OQ, [11]), i.e. the duration of the glottal open phase is calculated in two ways: the first one is the estimation of OQ through the inverse-filtered acoustic signal, in the same inversion paradigm as for the NAQ estimation, named OQ_A and the second is the articulatory values extracted from ElectroGlottoGraphic (EGG) measurement, that is a direct reference parameter, named OQ_{EGG}. We are thus able to compare (1) OQ_A to OQ_{EGG} in order to evaluate the inversion paradigm artefacts (2) NAQ to OQ_A in order to propose some objective criteria for the NAQ algorithm evaluation. The estimation of F0 is processed in the same way.

2. Spontaneous expressive speech

The choice of a corpus of authentic expressive speech recorded in lab rather than an acted one was made for several reasons. First, evidence from neurophysiology showed that acted emotion do not follow the same cortical mechanism as non-acted one [6], as they are not due to physiological changes. Moreover, as shown by Aubergé and Cathiard [2], acted amusement for instance can be discriminated from non-acted one, with a strong inter-judge effect. This implies that one cannot make sure that acted productions are identical to non acted emotional expressions, as the ability of an actor to reproduce exactly spontaneous emotional expressions cannot be evaluated in an objective way.

Secondly, acoustic analyses require a high-quality recording that can only be performed in lab condition [4], which implies to develop protocols for the induction of emotional states. In addition, the choice of such a method enables the control of phonetic and linguistic contents by the use of a command language that constraints the subjects' vocal expression. Eventually, it allows the collection on the same utterances of various emotional states, which can also be expected to carry various voice qualities.

Speech material for that study was thus extracted from an authentic but controlled expressive speech corpus recorded in a quiet room and mainly composed among others of monosyllabic words [3]. Emotional states were induced by subjects thanks to a Wizard-of-Oz scenario, Sound Teacher, implemented on a devoted platform, specially developed for building emotional scenarios (E-Wiz software). The Sound Teacher scenario imitates a voice recognition-driven software enabling the users to implicitly learn vowels from foreign languages. It aims at inducing first positive then negative emotional states in the subjects by manipulating their performances. The collected corpus consists in utterances of monosyllabic French color names ([ʀuʒ], [ʒon], [sabl], [vɛʀ], [bʀik]) chosen for the repartition of their vowels within the phonological space, as well as utterances of [paʒ sɥivɑ̃t]. Acoustic and EGG signals were recorded synchronously.

3. Voice source parameters

3.1. Acoustic analysis

Two speakers were selected on the basis of clear emotional and comparable productions. After the segmentation of interesting stimuli from the raw corpus, the phonetic labeling was performed by an expert. Numerous productions of those two speakers for words supposed to be monosyllabic revealed the presence of an unexpected schwa at their end (e.g. [ʒonə] instead of the expected [ʒon]), making those words disyllabic. Schwas were therefore also included in analyzes, as well as other vowels.

Acoustic analyses, implemented on Matlab routines, were carried out for every stimulus in the corpus. Fundamental frequency and intensity were estimated thanks to algorithms developed at ICP, and were used to calculate numerous distribution parameters: mean, standard deviation, jitter, shimmer, range, percentiles, as well as modeled f0 contours. Moreover, spectral analyses were implemented to calculate spectral slope, Hammarberg index and average long-term voiced spectrum on 9 frequency bands, as proposed in [13]. Eventually, duration of phonemes and syllables were calculated from the phonetic labeling.

3.2. Amplitude-based parameters of the glottal flow

Amplitude-based parameters have been suggested to provide a more robust method than time-based parameters for analyzing voice quality. The most widely used among them is the Normalized Amplitude Quotient proposed by Alku et al [1]. NAQ can be considered as a normalization of the "declination time", defined by Fant [7] as $NAQ=\frac{UP}{EE}\times F0$,

where UP is the peak-to-peak amplitude of the glottal flow, -EE is the value of the negative peak of the glottal flow derivative and F0 the fundamental frequency. Automatic calculation of the normalized Amplitude Quotient was performed thanks to an algorithm developed by Parham Mokhtari at ATR, Japan, in the frame of the JST/CREST Expressive Speech Processing Project. This algorithm performs a calculation of NAQ from speech signal on automatically detected syllabic reliability centers. This enables a fully automated extraction of NAQ values, thus providing a measurement of voice quality on unlabelled spontaneous speech [12].

Gobl and Ní Chasaide [9] have proposed to extend amplitude-based parameters to the estimation of time-based parameters. Therefore, the open phase of the glottal pulse can be estimated by: $T_{1A}=\frac{\pi}{2}\frac{UP}{EI}+\frac{UP}{EE}$, where EI is the value of the positive peak of the glottal flow derivative.

π.UP/2.EI is considered as an estimation of the glottal flow opening phase duration and UP/EE corresponds to the closing phase duration. Therefore, OQ is estimated by T_{1A}.F0. The same algorithm was also used to implement the calculation of Open Quotient from amplitude domain OQ_A. Moreover, the estimation of F0 performed by the algorithm at every detected reliability center was extracted in order to be compared to other estimations of pitch.

3.3. EGG parameters

Electroglottography is a measurement of impedance and gives information about the area of the vocal folds contact. $F0_{EGG}$ can be reliably estimated from EGG signal. Henrich [10] proposes an autocorrelation method between EGG signal and its derivative for the estimation of duration of the glottal pulse open phase $T1_{EGG}$ and the EGG Open Quotient (OQ_{EGG}.).

4. Results

4.1. Phonemic influence on NAQ

When calculated from unlabeled continuous speech, NAQ is available only on reliability centers, i.e. vocoids as defined by Mokhtari [12]. Therefore, locations of these reliability centers were also extracted and matched to the expert phonetic labeling of the corpus to ensure that detected segments are actual vocoids. Table 1 presents the repartition of reliability centers according to the phonemic labels. 68% of them are found in vowels, and 15% in sonorants. Except vowels, the nasal consonant [n] is often detected as a reliability center, and will hence be taken into account for further analyses.

Table 1: *Repartition (%) of the reliability centers according to phonemic labels.*

i	ɛ	a	o	u	ə	ɑ̃	n	others
9.4	11.6	14.7	7.3	8.8	3.0	13.2	8.3	23.7

Figure 1 shows the mean values and confidence interval of NAQ for each phoneme. NAQ ranges from 0.07 to 0.32, which has to be compared with Alku et al.'s [1] results obtained from five male speakers: pressed (0.08-0.11), modal (0.11-0.17) and breathy (0.23-0.35). Mean values of NAQ seem to be higher for higher oral vowels, however this tendency is not significant. The phoneme [ə] shows a higher NAQ. This trend is due to a clearly bimodal repartition of NAQ values. Speaker 1 adds [ə] on word endings with a high F0 and a high NAQ (0.28), which corresponds to a breathy voice. Speaker 2 produces schwas with a modal voice: NAQ values are about 0.12, as for [ɛ]. The choice of producing or not a final schwa seems to reveal a speaker-specific strategy related to speech-act expressive values. The nasal vowel [ɑ̃] shows NAQ values similar to high vowel ones. The nasal consonant [n] has NAQ values about 0.19, which can be interpreted as a breathy voice. All differences are significant except between [n] and [ə]. However, it seems unrealistic that the phoneme [n] in [ʒon] is always produced with a breathy voice, while the vowel [o] is not. This might be due to its final position, but high NAQ values are also measured when [n] is followed by a [ə]. A possible explanation is that nasality produces mainly low frequencies, thus attenuating higher frequencies and increasing the spectral slope. Both nasality and breathiness acoustically correspond to an increase in the spectral slope induced by supra-laryngeal settings for nasality and laryngeal settings for breathiness.

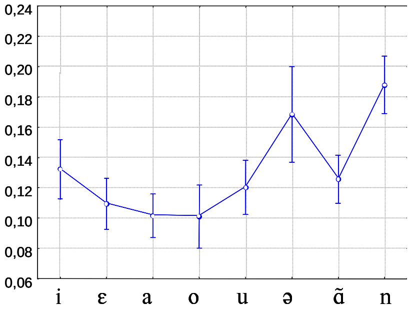

Figure 1: *Mean values and confidence interval p<0.01 of NAQ for each phoneme*

4.2. F0 estimations

Since most of amplitude-based parameters are normalized by the fundamental frequency, it implies that errors on its estimation also imply errors on the estimation of all the other parameters.

Figure 2 shows the fundamental frequency $F0_A$ estimated by the amplitude-based parameters algorithm plotted versus $F0_{EGG}$, i.e. fundamental frequency values obtained from the EGG signal. The correlation between both measurements of pitch is $r^2=0.64$. This should be compared to F0 values calculated by a prosodic editor EdiProso developed at ICP (threshold-based detection of signal cancellation points) for which the correlation with $F0_{EGG}$ reaches a value of 0.79.

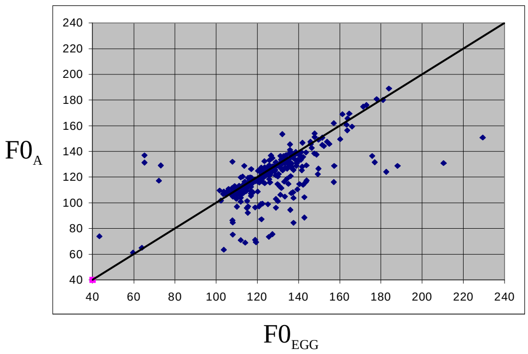

Figure 2: *$F0_A$ plotted vs. $F0_{EGG}$*

It appears from the comparison of F0 measurements that the fundamental frequency values used for the normalization of amplitude-based parameters tend to be underestimated. Therefore, normalized amplitude-based parameters values will also tend to be underestimated.

In our corpus, and for the two selected male subjects, pitch values grouped by phoneme reveal significantly higher pitch values for [ə] realized by Speaker 1, high pitch coupled to a high NAQ. Note that the two speakers nearly show the same proportion of added [ə]: Speaker 1 adds a schwa at the end of 36.8% of stimuli, against 42.9% for Speaker 2. However, the two speakers reveal different strategies in using schwa, showing that even in such a constrained protocol, speakers use different expressive strategies.

The high values of NAQ going with the high values of $F0_{EGG}$ brought us to calculate the correlation between NAQ and $F0_{EGG}$ which is $r^2=0.33$. If not null, this is a low value: NAQ and F0 are two independent parameters. We are more surprised by the fact that, in our data, the correlation between AQ (without normalization) and $F0_{EGG}$ is $r^2=0.08$, which is less than with normalization.

4.3. OQ_A vs. OQ_{EGG}

OQ is the duration of the open phase normalized by F0, i.e. the sum of the opening phase and the closing phase. Therefore, OQ_A, amplitude-based estimation of OQ, and NAQ, which is related to the closing phase [1], should be partly correlated. In our data, the correlation is $r^2=0.93$. This high correlation seems to indicate that the closing phase is sufficient to explain most of the open quotient variance, the asymmetry between the opening phase of the glottis and the closing phase being less important.

The correlation between OQ_A and $F0_{EGG}$ is $r^2=0.28$. Pitch values cannot explain the variation of the open phase duration, which seems to be clearly independent of other prosodic parameter.

Open quotient values measured from EGG signal OQ_{EGG} show no correlation with F0. These results may be compared to those obtained by Henrich [10] for singing voice. She compared F0 and OQ for different laryngeal mechanisms, and found a correlation between F0 and OQ in singers using laryngeal mechanism II, but not for mechanism I which is the most frequently used by male subjects in spoken sentences.

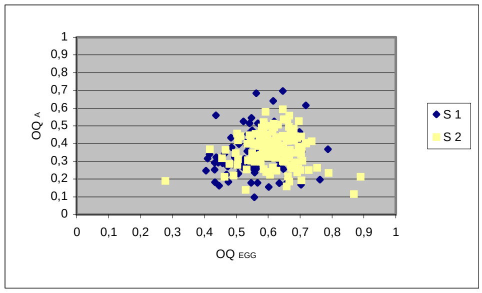

Figure 3: *OQ_A plotted vs. OQ_{EGG}*

Figure 3 shows values of amplitude-based Open Quotient OQ_A plotted against OQ_{EGG}, i.e. values extracted from the EGG signal. OQ_A values appear to be lower than OQ_{EGG} values. This pattern can be partly explained by the fact that $F0_A$ is underestimated. However, a similar pattern is also observable when $T1_A$ is plotted against $T1_{EGG}$, despite F0 values do no affect the calculation of $T1_A$ and $T1_{EGG}$. Indeed, $T1_A$ values are always smaller than $T1_{EGG}$ values, and no correlation is found either between $T1_A$ and $T1_{EGG}$, or between OQ_A and OQ_{EGG}. Even when considering each phoneme separately, the correlation coefficient was always significantly different for the two speakers.

These results are not consistent with Gobl and Ní Chasaide's [9] findings that OQ_A and OQ calculated from time domain are correlated with a coefficient of 0.76. In our study, we compared two different ways of estimating glottal parameter, one from inverse filtering, the other from impedance measurement of the glottis, when Gobl and Ni Chasaide were comparing two measures extracted from the output of inverse filtering.

5. Discussion

The first point to be underlined is that, though highly correlated in our corpus, NAQ and OQ_A characterize quite different phenomena, supposed to evaluate respectively the part of the closing phase and that of the open phase of the glottis. The energy of the glottal source is more produced during the vocal fold contact, that is the closing phase, than when the glottis is open. Therefore, NAQ estimations from the speech signal may be more reliable than OQ_A estimations. Moreover, they are calculated differently, since the calculation of OQ_A requires the estimation of one more parameter than the calculation of NAQ, namely EI, the maximum positive peak of the glottal flow.

It is rather surprising however to see that OQ_A is so weakly correlated to OQ_{EGG}, when Gobl and Ní Chasaide [9] found a high correlation between OQ values calculated from time and amplitude domains. One possible explanation would be that the inverse filter calculation used for estimating the glottal flow was not the better adapted one. Indeed, the amplitude-based parameters were calculated automatically, without any specific speaker adaptation, when Gobl and Ní Chasaide's results were obtained after performing an expert formant matching. Indeed, the evaluation of the glottal flow estimated by inverse filtering remains a specific problem: no known method provides a direct measurement of the glottal flow, so the survey by an expert appears as the better way to ensure correct inverse filtering.

In spite of our attempts, we are unable to link the articulatory-EGG measurements of vocal folds movements with the acoustic amplitude-based estimations of the glottal flow. However, glottal flow characteristics have been shown to influence perceptive emotional judgment [9], the NAQ parameter being related to the degree of care in the voice as shown by Campbell [5]. Obviously, NAQ estimation is a parameter extracted from the speech signal that carries information on the voice quality.

6. Conclusion

From a corpus of authentic expressive speech recorded in lab conditions [3], we have compared voice quality parameters obtained from amplitude domain by inverse-filtering the acoustic signal to direct measurements extracted from the synchronously recorded EGG signal. Amplitude-based parameters were calculated thanks to an algorithm performing automatic NAQ calculation from unlabeled acoustic signal [12]. This algorithm was also applied to the calculation of OQ_A from amplitude domain [9].

The results have shown a phoneme effect on NAQ, though with a different pattern for the two tested speakers. The use of NAQ as a prosodic parameter should be normalized by the phonemic factors. Moreover, calculated NAQ values on nasal [n] segments, frequently detected as vocoids, revealed to be overestimated. This can perhaps be related to the one-to-many problem of inversion, namely vocal tract settings linked with nasality and vocal folds control for breathiness producing similar acoustic effects.

However, it must be pointed out that, though these results could question the validity of direct dynamic measurements of NAQ, it does not concern the relevance of global/static estimations of NAQ when calculated on very large, implicitly phonetically balanced, corpora (e.g. [5]).

Comparison of $F0_A$ values estimated by the NAQ calculation algorithm to those extracted from EGG signal, $F0_{EGG}$, showed an underestimation of $F0_A$, due to pass on the normalized parameters OQ_A and NAQ.

Eventually, comparing OQ_A values to OQ_{EGG} did not show any correlation between those parameters, yet supposed to both estimate Open Quotient. This absence of correlation suggests however an inadequacy of the glottal flow estimated by inverse filtering and used in the calculation of OQ_A. An interesting prospect would be therefore to perform a speaker adaptation prior to the estimation of glottal flow.

7. Acknowledgments

This work was held within the Japan Science and Technology Agency as part of the CREST/Expressive Speech Project, directed by N. Campbell. We are deeply grateful to Parham Mokhtari and Nick Campbell for their technical solutions and fruitful advice.

8. References

[1] Alku, P.; Bäckström, T.; Vilkman, E., 2002. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustic Society of America*, 112 (2), 701-710.

[2] Aubergé, V.; Cathiard, M., 2003. Can we hear the prosody of smile ? Special issue *Emotional Speech*, *Speech Communication Review* 40.

[3] Aubergé, V; Audibert, N; Rilliard, A, 2003. Why and how to control emotional speech corpora. *8th European Conference on Speech Communication and Technology*, 185-188.

[4] Campbell, N., 2000. Databases of Emotional Speech. *ISCA Workshop on Speech and Emotions*, Newcastle, Northern Ireland, 34-38.

[5] Campbell, N.; Mokhtari, P., 2003. Voice Quality: the 4th Prosodic Dimension. *15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2417-2420.

[6] Damasio, A. R., 1994. *Descartes' error. Emotion, reason, and the human brain.* A. Grosset/ Putnam Books.

[7] Fant, G., 1997. The voice source in connected speech. *Speech Communication Review* 22, 125-139.

[8] Gobl, C.; Ní Chasaide, A., 2000. Testing affective correlates of voice quality through analysis and resynthesis. *ISCA Workshop on Speech and Emotions*, Newcastle, Northern Ireland, 178-183.

[9] Gobl, C.; Ní Chasaide, A., 2003. Amplitude-based source parameters for measuring voice quality. *ISCA Workshop on Voice Quality VOQUAL'03*, 151-156.

[10] Henrich, N. ; d' Alessandro, C.; Castellengo, M. ; Doval, B., 2000, Mesures électroglottographiques de quotient d' ouverture en voix parlée et chantée. *XXIIIèmes Journées d' Etude sur la Parole*, Aussois, France.

[11] Laver, J., 1980. *The phonetic description of voice quality.* Cambridge University Press, Cambridge.

[12] Mokhtari, P.; Campbell, N., 2002. Automatic Detection of Acoustic Centres of Reliability for Tagging Paralinguistic Information in Expressive Speech. *3rd International Conference on Language Evaluation and Resources*, Las Palmas, Spain, 2015-2018.

[13] Scherer, K. R.; Johnston, T.; Klasmeyer, G., 2003. Vocal Expression of Emotion. In R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Eds). *Handbook of Affective Sciences*, 433-456.