Fine-grain voice strength estimation from vowel spectral cues

Jean-Sylvain Liénard, Claude Barras

InterSpeech 2013 Lyon France
Communication #1076
Final version

Fine-grain voice strength estimation from vowel spectral cues

Jean-Sylvain Liénard, Claude Barras

LIMSI-CNRS, Univ. Paris Sud, 91403 Orsay, France
jean-sylvain.lienard@limsi.fr, claude.barras@limsi.fr

Abstract

This study investigates the possibility to recover the voice strength, i.e. the sound level produced by the speaker, from the signal recorded. The dataset consists of a set of isolated vowels (720 tokens) recorded in a situation where two interlocutors interacted orally at a distance comprised between 0.40 and 6 meters, in a furnished room. For each token, voice strength is measured at the intensity peak, and several sets of acoustic cues are extracted from the signal spectrum, after frequency weighting and intensity normalization. In the first phase, the tokens are grouped into increasing voice strength categories. Discriminant Analysis produces a classifier which takes into account all the signal dimensions implicitly coded in the set of cues. In the second phase, the cues of a new token are given to the classifier, which in turn produces its distances to the groups, providing the basis for estimating the unknown voice strength. The quality of the process is evaluated either in self-consistency mode or by cross-validation, i.e. by comparing the estimate with the value initially measured on the same token. The statistical margin of error is quite low, of the order of 3 dB, depending on the sets of cues used.

keywords: vocal effort, vocal intensity, voice quality, discriminant analysis

1. Introduction

It is common experience to recognize the voice strength adopted by a speaker, whatever the sound level actually received by the listener. An amplified weak voice remains perceived as a weak voice. This is true not only for the extreme modes such as shout or whisper, but also for the conversational speaking modes used everyday by any of us.

The objective of the present study is to demonstrate that, at least in the limited framework of an isolated vowel dataset, it is possible to recover the sound level emitted by the live speaker, by using acoustic cues extracted from the signal normalized in intensity.

1.1. State of the art

Our objective is a matter of several knowledge fields. The following references are not exhaustive. Relevant research work has been done on shouted speech intelligibility [1], on the role of the distance between interlocutors [2], on the effects of an excessive vocal effort [3][4], on the functioning and modeling of the vocal source [5][6][7], on the effect of the vocal effort on the acoustical structures of the speech signal [8][9], on the quality of the singing voice [10], on speech prosody [11], to mention just a few.

Most authors agree on the following considerations:

- speaking louder yields a frequency shift of the so-called "glottal formant", which is not to be confused with the modes of the vocal tract. This seems to be due to the shrinking of the open phase of the glottis. Usually located in the region of the first harmonics for a weak voice, the glottal formant may shift towards the higher harmonics in loud voice.
- the magnitude difference between the 1st harmonic H1 and the second H2 is strongly related to the open quotient Oq, but it also depends on the location of the 1st formant F1, i.e. on the vowel uttered.
- speaking louder or shouting yields a relative spectral emphasis of the high frequencies named spectral tilt. This feature is often estimated as the difference between H1 and the magnitude A3 of the 3rd formant.
- the fundamental F0 increases with the vocal effort; an order of magnitude is 15% per 6 dB intensity increase.
- the frequency F1 of the first formant, correlated with the speaker mouth aperture (jaw lowering), may increase of a quantity comprised between a few Hz to 200 Hz. The other formants do not appear to be systematically altered.

All of those effects are mixed with the other characteristics of the signal such as its phonemic and prosodic structures. Thus they are extremely difficult to extract from the signal without prior knowledge.

1.2. Conditions of the study

Sound intensity is the very object of our investigation. Thus it is essential to clearly distinguish the sound level produced by the speaker in live conditions, from the level of the same signal restituted by an electroacoustic device.

The concept of Vocal Effort being somewhat qualitative, we need an objective notion to represent it. We hypothesize that vocal intensity, measured in well defined conditions, is a good indicator for it. Vocal intensity should be expressed in dB SPL as in most of the above references, but in the present study we prefer the term "Voice Strength" because our data were not calibrated with a sound meter and thus cannot be taken as absolute measures, although they are reproducible throughout the database. Another reason is that we will use a spectral weighting, which greatly changes the figures representing sound intensity.

The voice strength shades occur and are recognized in everyday situations; instead of well controlled laboratory speech we will use data recorded in ordinary oral communication situations and exclude for the time being the modes of shouting and whispering which are less frequently used.

As many unknown factors may contribute to the determination of voice strength, our analysis method will not take for granted that voice strength varies linearly with any of the cues extracted from the signal.

2. Data and method

2.1. Data

The databases available in the speech community do not guarantee a rigorous constancy of the recording conditions, which is necessary to investigate the voice strength problem. Thus we used ours, a database named CORENC, formerly described and used in ref [8]. It is composed of 12 French vowels (9 orals, 3 nasals) uttered in isolation by 13 speakers (6 males, 7 females, aged 19-89). Three degrees of vocal effort noted "p", "n" and ""l were elicited by varying the distance between speaker and operator (0.4 m, 1.50 m and 6 m); however the subjects did not always produce the expected degree of vocal effort, so that the actual voice intensity was only loosely related to the distance condition. Consequently the latter was not taken into account in the present study. The recording took place in a quiet furnished room in 3 sessions, the 3rd of which was done six months after the first. Each speaker participated in a number of sessions comprised between 1 and 3. The speaker was seated, his/her mouth 30 cm away from the omnidirectional microphone. The task was to repeat interactively a short sentence and the set of vowels uttered by the operator. No calibration of the absolute sound level was done; however the recorder settings were kept identical during all of the 3 sessions, allowing a consistent measurement throughout the recordings.

2.2. Signal analysis

Each of the 720 tokens is represented by two entities. The first one is the peak intensity **a** of the token. The 2nd entity is a set of acoustic cues measured on the spectrum after its normalization to an arbitrary value that eliminates the voice strength direct information. Then a Discriminant Analysis classifier is trained on the data, in order to give an estimate **â** of the voice strength from the cues alone. The performance criterion is the standard deviation of the difference **<â-a>**, computed on the whole database. This margin of error is expressed in dB like the voice strength. The weaker the error, the more relevant the set of cues. The study has been done with the Praat software [12].

2.2.1. Signal intensity and frame selection

Signal intensity is computed on a 50 ms sliding Gaussian window. The frame is selected at the time where intensity reaches its peak value. This maximum intensity **ax** expressed in dB is considered as the voice strength of the token.

A second measure **ap** is performed in the same conditions, except that the signal is high pass filtered according to a function close to the A-weighting curve. The latter is used in sound level metering to take into account the reduced sensitivity of the human ear in the low frequencies, especially at low to medium level. Its origin goes back to the early work on loudness [13]. The attenuation goes approximately from 0 dB at 1000 Hz to 3 dB at 500 Hz and 23 dB at 63 Hz. The intensity of the A-weighted signals is expressed in dBA.

In the following, **axc** and **apc** represent the estimates of the measured values **ax** and **ap**.

2.2.2. Frame analysis

The signals are analyzed by a filter bank of 18 1-Bark wide filters. The outputs are integrated on a 50 ms Gaussian window. Then the frame spectrum is normalized with respect to its maximum.

According to the indices under study one can adopt the minimal resolution (some global cues may be computed from 18 channels separated by 1 Bark) or a better resolution by interpolation (when looking for spectral details, or for a visual representation). The fundamental frequency F0 is a basic cue, generally retained among the other acoustic cues. The Bark spectral modules are treated in Praat as Ltas objects (Long-term average spectrum).

2.2.3. Sets of acoustic cues

Many sets of acoustic cues can be extracted from the Ltas. In the present study the following sets were selected:

- F0 and Bark channels (b1 to b18): amplitudes in dB after normalization of the maximum to an arbitrary 50 dB.
- F0 and slopes: overall slopes (p5k is the amplitude difference between the high and low parts of the spectrum, trn and trn0 both represent the slope of the Ltas trend line); slopes in the low-part of the Ltas (p1k is the amplitude difference between the two halves of the 1-9 Bark band, d0 is the difference at the origin between the trend line and the Ltas); local slopes bij between adjacent channels i and j.
- F0 and voice perturbations: phonation irregularities such as jitter (jit), shimmer (shi), mean F0 (F0m), maximum autocorrelation (cor), harmonicity (hnr), all estimated on the whole token duration.
- F0 and Centres of Gravity taken on the whole Ltas (fg0, ag0), as well as on halves and quarters (fgi, agi, order i from 1 to 6). These cues must not be confused with the physical centres of gravity defined in linear frequency and magnitude scales. For those measures an intensity threshold is set at -40 dB from the maximum.
- F0 and 10 harmonics: amplitudes in dB of the first 10 harmonics, measured from the maximum on the interpolated Ltas at frequencies multiples of F0.

2.3. Discriminant Analysis

In order to establish a potentially non-linear matching between cues and voice strength, the tokens are grouped into a set of voice strength categories, according to a 3 dB step. Each category is given a label representing the mean voice strength of its component tokens. Then labels and cues are treated by Discriminant Analysis which, in the training phase, determines a set of orthogonal axes and their eigenvectors by maximizing the ratio between intergroup and intragroup variances. The eigenvectors are arranged into a set of functions called classifier (or discriminant). In the decision phase, the cues of a new token are presented to the classifier, which computes its distances to the group centers and proposes a set of candidates in terms of probabilities. The voice strength estimate is recovered by interpolating between the closest candidates.

2.3.1. Self-consistency and cross-validation modes

In self-consistency mode all the tokens are used to train the classifier; the same data are used as test. This mode is equivalent to a classification. It is legitimate if the goal is to investigate the structures of a closed dataset, but it is biased if the goal is to use the classifier with new, unseen data. We check this bias by performing a cross-validation. As the corpus

consists of 20 series of 36 tokens, a set of 19 series serves for training and the remaining series is used as test. The process is applied again 19 times within a circular permutation, and the error margin is computed as before on the 720 estimates.

3. Results

3.1. Sets of cues and A-weighting

3.1.1. F0 and 18 Bark channels, A-weighted signals

This set gives good results: error margins amounts to 2.94 dBA in classification, 3.25 dBA in cross-validation. This case will be used to illustrate the classification processing.

The voice strength dimension is quantified into 12 groups covering the 50 to 85 dBA interval in 3 dBA steps. The groups are labeled "05" to "16". Fig 1 (top left) displays the dispersion areas (at 1 sigma) of the groups in the plane of the 2 first classification functions. This shows the continuity and overlap of the groups, as well as the non-linearity that mainly affects the low end of the voice strength scale.

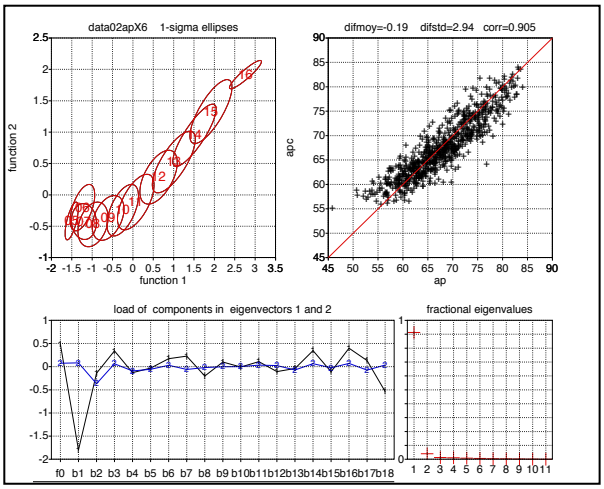

Figure 1: *Dispersion ellipses in the main factorial plane (top left); profiles of the first two eigenvectors (bottom left); fractional eigenvalues (bottom right); voice strength estimate vs. measure distribution (top right).*

Fig 1 (bottom) displays the components of the two main eigenvectors in terms of the 19 cues, and the fractional eigenvalues. The main eigenvector explains 91% of the variance and the second one only 4%. The first vector is dominated by the amplitude b1 of the 1st Bark channel. The second is dominated by b2. The contribution of the other cues is less prominent, at least in the main plane.

Fig 1 (top right) shows that, despite the quantification and the non-linearities, the (ap,apc) distribution is well concentrated along the diagonal. A slight asymmetry appears at the low end, meaning that this set of cues overestimates the voice strength of the weakest tokens.

3.1.2. Same analysis with unweighted signals

Fig 2 shows the same experiment on the unweighted signals. The results are almost equivalent in terms of error margin: 2.86 dB in self-consistency, 3.15 in cross-validation. However the non-linearity of the dispersion areas has augmented. More importantly, the (ax,axc) distribution is much farther from the diagonal, the correlation coefficient is poorer, and the number of groups built by the Discriminant Analysis is lower: we now have 8 groups instead of 12. Besides, the voice strength dynamics is poorer (some 25 dB, instead of 35), while the error margin remains comparable, about 3dB.

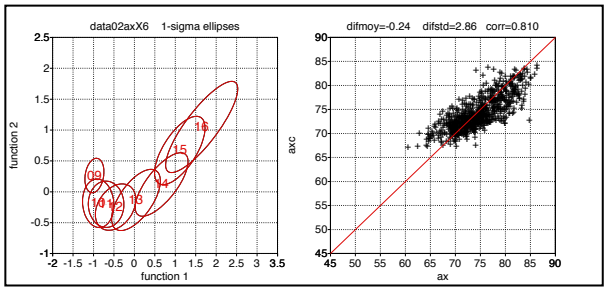

Figure 2: *Dispersion areas and (ax,axc) distribution in the absence of the A-weighting of the signals.*

All this indicates that the A-weighting favors to a large extent the coding of the voice strength, especially for the normal to low voices. In the rest of the study only the A-weighted signals will be taken into consideration.

3.1.3. F0 and 4 slopes: p1k, p5k, trn, d0

This set provides a poorer margin of error: 3.57 dBA, 3.75 in cross-validation. However, as only 5 cues are used, they can be considered more relevant and more general than the above 18 cues, except b1 and b2. As in the previous experiments, the weight of F0 in the main eigenvectors is low.

3.1.4. F0 and 6 voice perturbation cues

For this set of cues the performance is poor (4.59 dBA, 4.75 in cross-validation) and the (ap,apc) distribution diverges from the diagonal. Actually, this result is much better than expected, considering that it does not contain any spectral information. The most significant cues seem to be cor, F0m and hnr.

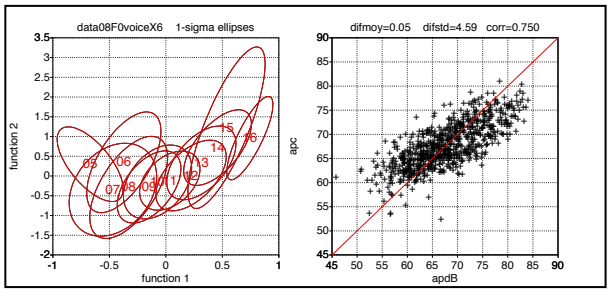

Figure 3: *Dispersion areas and (ap,apc) distribution for the F0+6 voice cues*

3.1.5. F0 and 6 cues from 3 centres of gravity

This set comprises 7 cues: F0 plus the amplitudes and frequencies of 3 centres of gravity of orders 0, 1 and 2. The results are poor (3.88 dBA, 4.05 in cross-validation) and the examination of the (ap,apc) distribution shows non linearity and a lack of efficiency in its lower part. Again, F0 seems to play a secondary part.

3.1.6. F0 and 10 harmonics

This set produces an acceptable performance (3.28 dBA, 3.52 in cross-validation), mainly due to F0 and h1 which play in opposite directions.

3.2. Compound sets

In this section we compose new sets by selecting some of the most significant cues found in the previous experiments.

3.2.1. 34-cue set

This large set includes the best cues found in each of the previous sets. The results are better than with any of the initial complete sets: 2.59 dBA, 3.03 in cross-validation. The (ap,apc) distribution is closer to the diagonal and more symmetrical than in any of the previous experiments. The crescent shape observed in the main plane has no negative consequence on the performance, because its projection on the main axis, which explains 85% of the intergroup variance, is perfectly regular. It suggests that different cues contribute to the high and low parts of the voice strength interval.

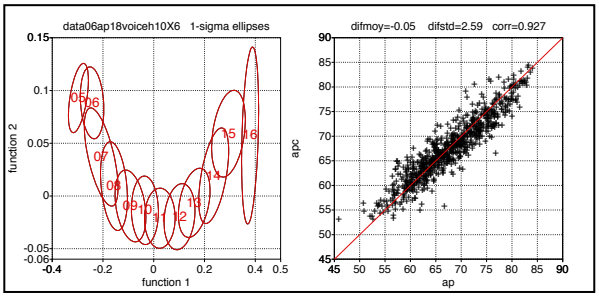

Figure 4: *Dispersion areas and (ap,apc) distribution for the compound set of 34 cues.*

3.2.2. Looking for the most important cues

Examining the main eigenvectors allows a ranking of the most important cues: p5k, d0, ag5, ag6, trn0, ag1. However, gathering them into a new, reduced compound set yields a margin of error that is not bad, but not much better than some of the primary sets: 3.37 dBA (3.49 in cross-validation). Thus in the present state of the study one can only propose the following qualitative observations:

- the global slope of the A-weighted Bark Ltas, taken from the medium to the high frequency bands, seems to be the main underlying factor of voice strength estimation. It is reflected in the cues that appear frequently at the top of our cues lists, such as p5k, trn0, ag5 and ag6. It is obviously related to the spectral tilt evidenced in many previous studies.
- The slope in the lower part of the Ltas also plays an important role, which appears in cues such as d0, ag1, or b1 or h1 in previous experiments. It may be related to the H1-H2 or H1-A1 parameters found in other studies.
- F0, taken as an absolute cue, seems to play a secondary role. It could be rated differently if it were taken as a difference, for a given talker, between the instantaneous and the mean F0 values.

4. Discussion

The main point that emerges from the present study is that in conditions close to ordinary oral communication, the speech signal, even reduced to a mere spectrum taken in the central part of a vowel, contains enough information to characterize the voice strength within a very low margin of error.

The closeness of the figures obtained in self-consistency and in cross-validation indicates that the proposed processing does not depend much on the variations due to new data. However some variability due to the talker and phonetic content may remain in the results, which leaves the door open to further improvements.

The use of a frequency weighting close to the psychoacoustically grounded A sound metering norm, is amply justified by the results, because it yields more degrees in the estimation of the voice strength.

In vocal quality studies, vocal effort in the conversational range is usually qualified with 3 timbre qualities only: modal voice, loud and low voice. This particular aspect of timbre could be replaced by voice strength which, considering the 3 dBA margin of error and the 35 dBA interval, may provide a number of degrees closer to 10 than to 3.

The link between vocal effort and F0 has been widely demonstrated experimentally in the literature. Talking louder usually yields an irrepressible raise of F0, except for especially trained talkers. The reciprocal is not true: any talker can raise F0, up to about one octave, without augmenting his/her voice strength. This capability is used to transmit intonational information. Besides, at equal voice strength, F0 is strongly related to the talker's position in the {male, female, child} vocal categorization. As we choose from the start of the study not to use any prior information on the talker and context, the secondary role found for F0 simply confirms the non-reciprocity of the relation between F0 and voice strength.

5. Conclusions

The approach proposed, based on discriminant analysis and on a database with controlled recording conditions, has shown that voice strength could be deduced from an uncalibrated speech signal with surprising precision. The method may prove useful in other aspects of voice and speech research, for instance for investigating the relations between phonemic systems and speaker characteristics. Some developments may be envisioned in several fields of automatic processing. In synthesis, mastering the voice strength may help to produce more natural, situation-sensitive voices. In speech recognition, knowledge of the voice strength may contribute to reduce the variability of the linguistic speech structures. In speech diarization or speech scene analysis, differences in voice strength may contribute to separating the talkers voices.

6. Acknowledgements

The authors would like to thank the scientific committee of LIMSI-CNRS, as well as their colleagues Drs Christophe d'Alessandro, Philippe Boula de Mareüil, Albert Rilliard from LIMSI, and Pierre Divenyi from CCRMA, Stanford Univ.

7. References

[1] Rostolland, D., "Acoustic features of shouted voice", Acustica, vol 50, 118-125, 1982.

[2] Traumuller, H. and Eriksson, A., "Acoustic effects of variation in vocal effort by men, women and children", J. Acoust. Soc. Am. 107 (6), 3438-3451, 2000.

[3] Junqua, J.-C., "The Lombard reflex and its role on human listeners and automatic speech recognizers", J. Acoust. Soc. Am. 93, 510-524, 1992.

[4] Garnier, M., "Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal", doctoral thesis, Paris VI university, 2007.

[5] Fant, G., Liljencrants, J. and Lin, Q. "A four parameter model of glottal flow", STL-QPRS, 26(4):1-13, 1985.

[6] Doval, B., d'Alessandro, C. and Henrich, N. "The spectrum of glottal flow models", Acustica united with Acta Acustica, 92:1026-1046, 2006.

[7] Hanson, H. M., "Glottal characteristics of female speakers: acoustic correlates", J. Acoust. Soc. Am. 101 (1), 466-481, 1997.

[8] Liénard, J.-S. and Di Benedetto, M.G., "Effect of vocal effort on spectral properties of vowels", J. Acoust. Soc. Am. 106 (1), 411-422, 1999.

[9] Huber, J.E., Stathopoulos, E.T., Curione, G.M., Ash T.A. and Johnson, K., "Formants of children, women, and men: the effects of vocal intensity variation", J. Acoust. Soc. Am. 106 (3), 1532-1542, 1999.

[10] Henrich, N., d'Alessandro, C., Doval, B. and Castellengo, M., "Glottal open quotient in singing: measurement and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency", J. Acoust. Soc. Am. 117 (3), 1417-1430, 2005.

[11] d'Alessandro, C., "Voice source parameters and prosodic analysis", in S. Sudhoff et al [Eds] Methods in Empirical Prosody Research, 63-87, Walter de Gruyter, 2006.

[12] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" [computer program], version 5.3.32, retrieved 17 October 2012 from http://www.praat.org/.

[13] Fletcher, H. and Munson, W.A. "Loudness, its definition, measurement and calculation", J. Acoust. Soc. Am. 5, 82-108, 1933.