Singing voice quality assessment in professional singers using microphones and contact sensors

Antonella Castellana

Dipartimento di Energia, Politecnico di Torino, Italy

Ilaria Leocata

Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, Italy

Alessio Carullo

Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, Italy

Arianna Astolfi

Dipartimento di Energia, Politecnico di Torino, Italy

Summary

This study focuses on objective evaluations of singing voice quality obtained from the analysis of different singing tasks. Three devices were used to simultaneously acquire the performed tasks, thus allowing a comparison among the estimated parameters to be obtained. Fourteen professional singers took part in the experiment: they sang two Italian songs at comfortable tonality and performed an arpeggio using the vowel /a/, while standing in front of a sound level meter (SLM) and wearing two types of contact sensors, namely an Electret Condenser Microphone (ECM) and a Piezoelectric Contact Microphone (PM). They also read aloud an Italian phonetically balanced text. The singing voice quality was assessed by means of parameters related to pitch inaccuracy and singer's formant. In addition, Cepstral Peak Prominence Smoothed (CPPS) distributions were investigated in songs for the first time. The pitch inaccuracy estimation was comparable for the three devices: the overall mean of the pitch deviation between each contact microphone and SLM was equal to 1.9 Hz (standard error: 0.4 Hz), thus making ECM and PM as good as SLM for this estimation. Significant differences were found for the singer's formant evaluations, which were performed observing the Long-Term Average Spectra (LTAS) and calculating the Singing Power Ratio (SPR): since the signal acquired at the output of the contact sensors is affected by the physiological filtering (vocal folds – throat – skin), but not by the filtering effect of the vocal tract that affects microphones in air, the two contact sensors showed a higher spectral slope. However, a good correlation between SPR obtained from ECM and SLM proved contact microphones are able to highlight the presence of the singer's formant. Furthermore, evidence has been found that CPPS distributions shape from the three types of microphone indicate the degree of singing voice quality.

PACS no. 43.70.+i, 43.72.+q

1. Introduction

Voice is the primary mean of communication among people, thus having a great role in daily life. Voice is also the main tool that different professional categories use during their working hours, such as teachers, call center operators and singers. For these workers it is important to control their voice quality, in order to prevent voice disorders due to a misuse of the vocal apparatus. Singers also take care of the artistic nature of voice usage to have a great success during exhibitions and concerts. All the aspects of good singing can be evaluated perceptually by an experienced trained singer. However, the need of an objective evaluation has led to the spread of singing voice quality investigations by means of acoustic analyses of voice productions. Several objective measures

that are able to describe the singing performance have been studied in the existing literature.
One of the most important aspects of singing is the control of fundamental frequency. Previous studies investigated the pitch accuracy in trained and untrained singers through singing tasks acquired with a headworn microphone, while changing the external auditory feedback [1,2].
An extra formant, the so-called singers' formant, has been detected in singers' spectra: it adds clarity, projection, and timbral differentiation to the voice [3]. This formant presents a strong area of energy in the frequency range (2800÷3500) Hz, and the practical reason for its presence is that it permits the singer to be heard above the orchestra [4].
The Long-Term Average Spectrum, LTAS, which consists in averaged spectra over time, has also been widely used to investigate voice quality in both continuous speech [5] and sung passages [6]. The Singing Power Ratio (SPR), which is a measure derived from the spectrum of voice signals, is defined as the ratio of the peak intensities between the (2÷4) kHz and (0÷2) kHz frequency bands [7]. As such, it is able to account for the presence of the singer's formant. Previous works have found SPR values [8,9] and LTAS [10] significantly different between trained and untrained singers.
All the above-mentioned studies used microphones in air to acquire singing samples: such devices record the vocal signal at the output of the singer's lips that is affected by the vocal tract filtering, which is an aspect of great interests to be analyzed for the singing voice quality.
However, singers during training could take advantage of constant self-monitoring of vocal output in order to make frequent small corrections of the muscle activity for voice production. Contact sensors, which are attached to the singer's neck, are promising devices, since they sense the vibrations due to vocal folds activity and they have a negligible sensitivity to the background noise.
While microphones in air acquire the vocal signal that is modified by the vocal tract filtering, contact sensors acquire the voice source affected by the physiological filtering (vocal folds-throat-skin). As consequence, the signals acquired by the two types of microphones are different in both time and frequency domains [11]. Few studies have compared the output of an objective evaluation of singing voice quality performed using a microphone in air and a contact microphone [12, 13].
In the present study, different singing tasks were asked to professional singers while standing in front of a sound level meter and wearing two types of contact sensors, namely an Electret Condenser Microphone (ECM) and a Piezoelectric Contact Microphone (PM). The main aim was to compare the above-mentioned objective evaluations of singing voice quality obtained from the analyses of the signals acquired with the three devices. As a further investigation, Cepstral Peak Prominence Smoothed (CPPS) distributions, which have been already used to effectively discriminate between healthy and pathological voice [14], were investigated in songs for the first time.

2. Method

2.1. Subjects

Fourteen professional singers, 11 females and 3 males, participated in this experiment (age range: 21-56 years, mean: 36 years): 1 tenor (T), 1 bariton (Ba), 1 bass (B), 1 alto (A), 2 mezzo-sopranos (MS) and 8 sopranos (S).
All subjects provided the most relevant personal information, e.g. age, years of experience and their repertoire. They were native Italian singers.

2.2. Procedure

The experiment was performed in the anechoic chamber of Politecnico di Torino, where the A-weighted equivalent background noise level was 24.5 dB and the mid-frequency reverberation time (from 0.5 kHz to 2 kHz) was 0.11 s.
After an initial warm-up, the entire protocol was asked to the participants as follows:

- reading an Italian phonetic-balanced text;
- singing "Happy birthday" song (Song 1) in Italian language using a comfortable tonality;
- singing the Italian national anthem (Song 2) using a comfortable tonality;
- singing an arpeggio with two tempi, i.e. slow (40 bpm) and fast (120 bpm), and articulations, i.e. legato and staccato, using the vowel /a/.

The task order was randomized.

2.3. Equipment

Three different microphones were simultaneously used by each singer:

- a class-1 calibrated Sound Level Meter, *SLM*, (XL2, NTi Audio), which was placed on-

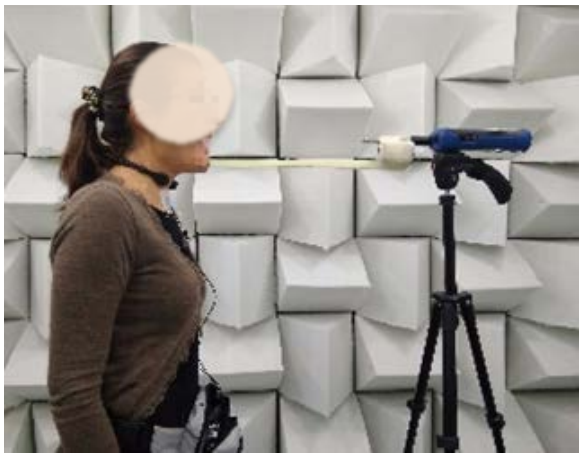

Figure 1. A subject while performing the experiment.

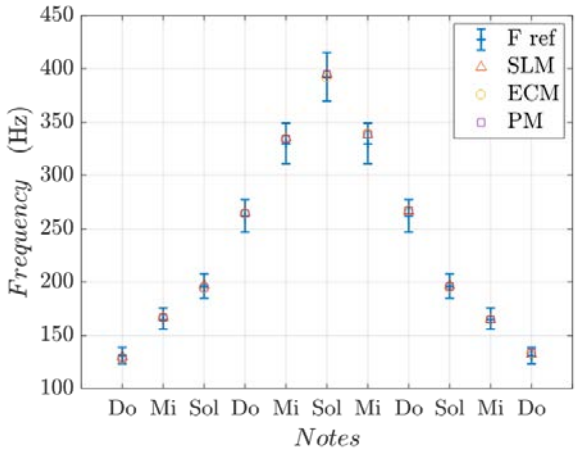

Figure 2. Pitch inaccuracy of the tenor evaluated with the three devices.

axis at a fixed distance of 30 cm from singer's lips by means of a spacer. The signals were saved into the internal memory using a sampling rate of 44100 Hz and 16 bit of resolution.

- an Electret Condenser Microphone, *ECM*, (AE38, Alan Electronics Gmbh, Dreieich, Germany) fixed at the jugular notch of the singer using a surgical band. It was connected to the portable recorder ROLAND R05 (Roland Corp., Milano, Italy) that samples the signal at a frequency of 44100 Hz using 16 bit of resolution;
- a piezoelectric contact microphone, *PM*, (HX-505-1-1, Shenzhen, China), which is embedded in a collar placed around the neck and connected to a smartphone (Samsung SM-G310HN). The recordings were performed using the Vocal Holter Rec (PR.O.VOICE, Turin, Italy) and saved into the internal memory of the smartphone using a rate of 22050 Hz and 16 bit of resolution.

Figure 1 shows a subject while performing the experiment.

2.4. Data processing

After each recording, data were stored in a PC.
Specifically designed MATLAB 2017A scripts were used to automatically align the signals acquired with the three devices and to estimate voice parameters that are related to singing voice quality.
The fundamental frequency was estimated in the 200 ms-middle part of each vowel /a/ of arpeggios in order to evaluate the pitch accuracy. Signals were oversampled at a frequency of 352.8 kHz to obtain a frequency resolution of about 0.5 Hz and the autocorrelation function was implemented.
In order to observe the prominence of the singer's formant, a proper script was implemented to evaluate the LTAS in each song. A 1024 samples analysis frame without overlap, weighted by means of a Hamming window, was used.
The signals from arpeggios were also used to evaluate the Singing Power Ratio, in order to quantify the prominence of the singer's formant. According to Omori *et al.* [6], a steady 4096-point portion of each vowel was selected, weighted by means of a Hanning window, and then processed using an FFT algorithm. The difference in dB between the two highest harmonics in the range 0 Hz÷2 kHz and (2÷4) kHz was calculated.
Furthermore, a proper algorithm for Cepstral Peak Prominence Smoothed (CPPS) distributions was used for readings, as described in Castellana *et al.* [15]. A modified version was implemented for songs: the peak search in the cepstrum domain was extended from (3.3÷16.7) ms to (1.7÷16.7) ms, thus considering frequencies between (60÷600) Hz in order to cover the higher frequency range of singers.

3. Results

3.1. Pitch inaccuracy

Figure 2 describes the results about the evaluation of pitch inaccuracy for the tenor while performing the arpeggio (articulation: legato; tempo: slow). The triangle, circle and square points indicate the fundamental frequency obtained from the signals acquired with SLM, ECM and PM, respectively. The center of the blue bars represents the reference

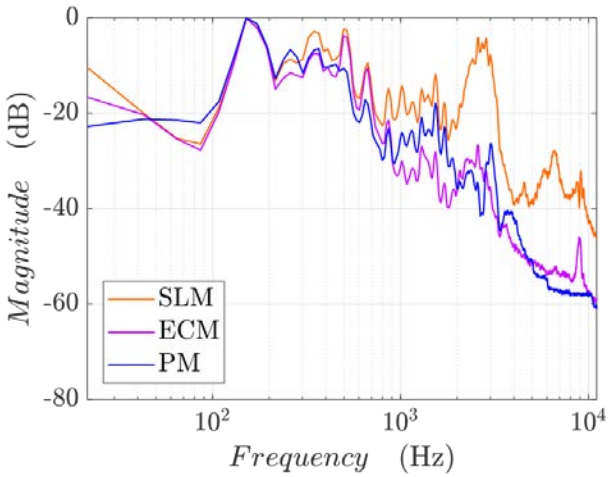

Figure 3. Overlapped LTAS of Song 1 performed by the bass for each measurement chain.

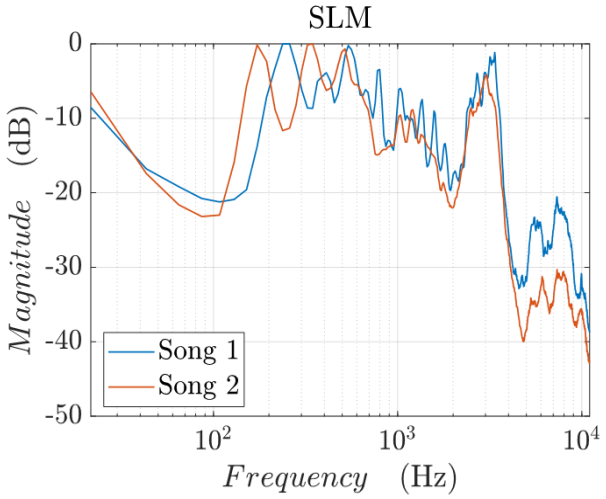

Figure 4. Overlapped LTAS of Song 1 and Song 2 performed by the bass and acquired with SLM.

note: the upper bound represents the highest semi-tone of the note, while the lower bound represents the lowest semi-tone. If the fundamental frequency is out of the bar, the singer is considered off-key. The performance is considered "crescente" or "calante" when the fundamental frequency falls into the upper or lower blue segment, respectively.

As shown in Figure 2, the fundamental frequencies estimated from the signals acquired with the three microphones are quite overlapped. Such a behavior has been observed for all the singers in the 4 types of arpeggios, i.e legato and slow, legato and fast, staccato and slow, staccato and fast. The overall mean of the pitch deviation between each contact microphone and the SLM is equal to 1.9 Hz (standard error: 0.4 Hz). This result highlights that the fundamental frequencies estimated from the contact devices and the microphone in air are comparable. These results corroborate and extend previous studies that have concerned with the comparison of fundamental frequencies estimated from voice signals acquired with microphones in air and accelerometers [12],[16].

3.2. Singer's Formant

3.2.1. Long-term Average Spectrum

Figure 3 shows the LTAS for each measurement chain obtained from Song 1 performed by the bass. Since the signal acquired at the output of the contact sensors is affected by the physiological filtering (vocal folds – throat – skin), but not by the filtering effect of the vocal tract that affects microphones in air, a higher LTAS slope for the two contact sensors was expected. However, in Figure 3 the LTAS of PM shows an unpredicted magnitude increase in the frequency range (1÷4) kHz. Previous studies highlighted that such a boost of energy content helps intelligibility, since PM is commonly used in very noisy environments, such as in the chockpit by helicopter drivers [17].

With respect to the singer's formant, the LTAS of SLM clearly shows a peak around 3 kHz, which is about 5 dB lower that the first harmonic peak. The LTAS of ECM shows a peak of energy at the same frequency, but about 25 dB less prominent that the SLM one. Differently from the previous LTAS, the LTAS of PM has the peak at 1.5 kHz that is more prominent than the one at 3 kHz, probably for its particular frequency response described above. As such, the PM is not suitable to highlight the presence of the singer's formant.

Figure 3 also underlines the frequency content in the acquired signals: the LTAS of the SLM shows frequency content up to 10 kHz; for the two contact sensors, instead, a lower frequency content is noticeable: about 3.5 kHz for ECM and about 5 kHz for PM.

Furthermore, since singers performed two songs with two different extensions, the authors were able to verify that the singer's formant does not depend on what he\she is singing. Figure 4 shows overlapped LTAS of Song 1 and Song 2 performed by the bass and acquired with SLM. Even if the singer sang the two pieces at different tonalities (see the first peaks in LTAS), the peak of the singer's formant corresponds in the two LTAS. The same outcome was also obtained for ECM and PM.

Table I. Mean SPR obtained for each singer in all the arpeggios (±standard deviation) for the three microphones; overall mean, OM, among the singers (SE=standard error); Pearson coefficient, P, between the obtained data.

	SPR_{ECM}	SPR_{SLM}	SPR_{PM}
B	-35.6(4.2)	-17.5(6.5)	-30.3(6.8)
Ba	-30.3(4.5)	-17.0(5.4)	-42.3(8.7)
T	-38.6(7.4)	-21.5(5.2)	-42.4(7.5)
A	-31.2(4.8)	-12.7(3.3)	-34.9(3.0)
MS	-46.2(8.7)	-20.1(6.3)	-34.0(7.7)
MS	-37.2(7.1)	-14.1(9.5)	-25.9(6.6)
S	-37.1(6.0)	-22.9(5.2)	-29.9(3.6)
S	-39.9(10.3)	-27.9(8.0)	-43.6(7.2)
S	-36.8(6.7)	-20.3(4.7)	-38.7(6.2)
S	-38.3(7.5)	-22.3(4.6)	-36.8(3.4)
S	-35.9(13.5)	-25.2(9.4)	-41.5(5.0)
S	-43.9(6.4)	-15.9(4.6)	-34.2(5.4)
S	-51.7(8.9)	-29.3(7.5)	-33.1(8.1)
S	-52.2(5.6)	-30.1(2.9)	-37.8(7.6)
OM(SE)	***-39.6(6.6)***	***-21.2(5.5)***	***-36.1(1.4)***
$P_{ECM-SLM}$	0.64		
P_{SLM-PM}		0.38	

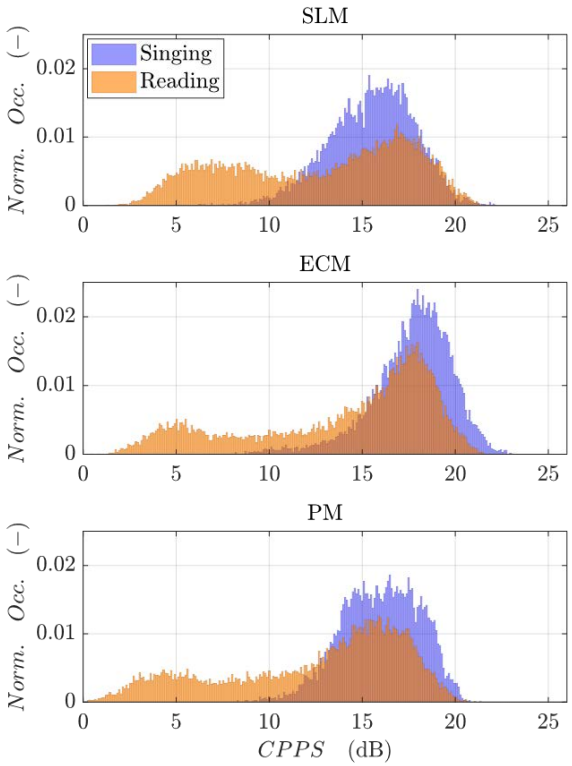

Figure 5. Overlapped CPPS distributions of singing (blue) and reading (orange) tasks obtained from a tenor. Vocal signals were acquired with the three devices.

3.2.2. Singing Power Ratio

Table I reports the mean values of SPR obtained for each singer in all the arpeggios for the three microphones. SPR values obtained for SLM corroborate previous works that have evaluated the parameter in performances of professional singers [8].

The overall means show that ECM and PM have similar SPR values and both have SPR values about 15 dB higher than SLM ones. Such a result confirms the evidences described in the previous paragraph.

In order to explore how different are the behaviours of the three microphones with this parameter, the Pearson coefficient was calculated between the lists of data of SPR from ECM, SPR_{ECM}, from SLM, SPR_{SLM}, and from PM, SPR_{PM}. Table I shows a good correlation of 0.64 between SPR_{ECM} and SPR_{SLM}, while a poor correlation of 0.38 has been found between SPR_{SLM} and SPR_{PM}.

The agreement between SPR_{SLM} and SPR_{ECM} could be explained by Sundberg's studies, which stated that "the acoustic situation that explains the singer's formant is that the larynx tube serves as an autonomous resonator with a resonance frequency in the vicinity of 3 kHz, that is not much influenced by the rest of the vocal tract" [18]. As consequence, both SLM and ECM can acquire singing signals where the singer's formant is observable. Regarding the difference between SPR_{SLM} and SPR_{ECM}, the evidences about LTAS of PM in paragraph 3.2.1 also affect the SPR values.

3.3. Singing voice quality using cepstral analysis

Figure 5 shows the overlapped CPPS distributions obtained for the reading and singing tasks (Song 1) of a tenor, which were acquired with the three devices. The orange distribution, which is related to the reading task, has a bimodal shape. Such an evidence is related to the simultaneous presence in continuous speech of phonemes with a regular spectrum (e.g. vowels), others that produce irregular spectra (e.g. consonants) and prosody, as descried in Castellana [17]. The blue distribution, which is related to the song, has a single-mode shape and it is overlapped to the highest CPPS values of the reading distribution. Such differences in distribution shapes are related to a better control of vocal folds during singing performance that

provides phonemes with a regular spectrum. Such a behavior can be observed for all the three devices.

4. Conclusions

The present study deals with the comparison of different objective evaluations of singing voice quality obtained from the analyses of singing tasks acquired with three devices. Fourteen professional singers were asked to sing two Italian songs at comfortable tonality and to perform an arpeggio using the vowel /a/, while simultaneously standing in front of a sound level meter (SLM) and wearing two types of contact sensors, namely an Electret Condenser Microphone (ECM) and a Piezoelectric Contact Microphone (PM). They also read aloud an Italian phonetically balanced text.

The three microphones are comparable for the pitch inaccuracy evaluation, but, as expected, significant differences were found for the spectral evaluations, i.e, LTAS and SPR. However, the good correlation found between SPR_{ECM} and SPR_{SLM} lead to preliminary conclusions that contact microphones could highlight the presence of the singer's formant. Furthermore, the shape of CPPS distributions from the three types of microphone could indicate the degree of singing voice quality.

Future studies should deepen the objective evaluation of singing voice quality using contact microphones by comparing outcomes from trained and untrained singers and by using larger database.

References

[1] P. Bottalico, S. Graetzer, and E.J. Hunter: Effect of Training and Level of External Auditory Feedback on the Singing Voice: Pitch Inaccuracy. *Journal of Voice, 31*(1) (2017) 122-129.

[2] C. Watts, J. Murphy, and K. Barnes-Burroughs: Pitch matching accuracy of trained singers, untrained subjects with talented singing voices, and untrained subjects with nontalented singing voices in conditions of varying feedback. *Journal of Voice*, *17*(2) (2003) 185-194.

[3] J. Sundberg: The science of the singing voice. DeKalb, IL: Northern Illinois University Press; 1987

[4] C. Watts, K. Barnes-Burroughs, J. Estis, and D. Blanton: The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers. *Journal of voice*, *20*(1) (2006) 82-88.

[5] A. Löfqvist, and B. Mandersson: Long-time average spectrum of speech and voice analysis. *Folia Phoniatrica et Logopaedica*, *39*(5) (1987) 221-229.

[6] T.F. Cleveland, J. Sundberg, and R.E. Stone: Long-term-average spectrum characteristics of country singers during speaking and singing. *Journal of voice*, *15*(1) (2001) 54-60.

[7] J. Sundberg: Level and center frequency of the singer's formant. *Journal of voice*, *15*(2) (2001) 176-186

[8] K. Omori, K. Ashutosh, L. Carroll, W. Riley, and S. Blaugrund: Singing power ratio: quantitative evaluation of singing voice quality. *Journal of Voice,* 10 (1996) 228–235.

[9] M. Usha, Y.V. Geetha, and Y.S. Darshan: Objective identification of prepubertal female singers and non-singers by singing power ratio using matlab. *Journal of Voice*, *31*(2) (2017) 157-160.

[10] V.M.O. Barrichelo, R. J. Heuer, C. M. Dean, and R.T. Sataloff: Comparison of singer's formant, speaker's ring, and LTA spectrum among classical singers and untrained normal speakers. *Journal of voice*, *15*(3) (2001) 344-350.

[11] A. Carullo, A. Astolfi, A. Castellana, G.E. Puglisi, F. Casassa, and L. Pavese: Performance comparison of different contact microphones used for voice monitoring, in Proceedings of the International Congress on Sound and Vibration 22, Florence, July, 12-16 2015.

[12] R.F. Coleman: Comparison of microphone and neck-mounted accelerometer monitoring of the performing voice. *Journal of Voice,* 2(3) (1988) 200-205.

[13] A. Lamarche, and S. Ternström: An exploration of skin acceleration level as a measure of phonatory function in singing. *Journal of Voice*, *22*(1) (2008) 10-22.

[14] A. Castellana, A. Carullo, S. Corbellini, A. Astolfi, M. Spadola Bisetti, and J. Colombini: Cepstral Peak Prominence Smoothed distribution as discriminator of vocal health in sustained vowel. In Proc. IEEE I2MTC, Torino, Italy, May 22-25 (2017) 552-557.

[15] A. Castellana, A. Carullo, S. Corbellini, and A. Astolfi: Discriminating pathological voice from healthy voice using Cepstral Peak Prominence Smoothed distribution in sustained vowel, *IEEE Transactions on Instrumentation and Measurement*, 67(3) (2018) 646-654.

[16] D. D. Mehta, J. H. Van Stan, and R. E. Hillman, "Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 4, pp. 659–668, Apr. 2016.

[17] A. Castellana: "Towards vocal-behaviour and vocal-health assessment using distributions of acoustic parameters," *Doctoral Thesis*, Politecnico di Torino (2018).

[18] J. Sundberg: Articulatory interpretation of the "singing formant."*J Acoust Soc Am.*, 55 (1974) 838-844.