

A new paradigm of effective communication based on voice shapes

Original

A new paradigm of effective communication based on voice shapes / Carullo, A.; Anibaldi, A.; Astolfi, A.; Atzori, A.; Cennamo, V.; Zito, G.. - ELETTRONICO. - 2019-:(2019), pp. 7781-7788. (Intervento presentato al convegno 23rd International Congress on Acoustics: Integrating 4th EAA Euroregion, ICA 2019 tenutosi a Aachen , Germany nel 9-13 September 2019) [10.18154/RWTH-CONV-238940].

Availability:

This version is available at: 11583/2941152 since: 2021-11-29T10:48:39Z

Publisher:

International Commission for Acoustics (ICA)

Published

DOI:10.18154/RWTH-CONV-238940

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

A New Paradigm of Effective Communication based on Voice Shapes

Alessio CARULLO¹; Adriano ANIBALDI²; Arianna ASTOLFI³; Alessio ATZORI¹;
Viviana CENNAMO¹; Giovanni ZITO²

¹ Politecnico di Torino - Electronics and Telecommunications Department, Italy

² Interago Academy, Rome, Italy

³ Politecnico di Torino - Department of Energy, Italy

ABSTRACT

A new relational paradigm is proposed that is based on voice shapes, which represent the speech style used to establish an effective communication. The functional voice shapes are: “rounded”, which means that a colloquial and empath communication is established; “triangular”, which means transmitting energy, joy and interest; “squared”, which highlights competence and solidity. The dysfunctional shapes are: “flat”, a monotone style that does not capture the listener attention; “spiky”, which is an aggressive style that transmits anger or blame towards the listener. An attempt has been made to match the voice shapes to acoustic features of the vocal signal, starting from parameters extracted from the recordings of 12 actors that have reproduced the voice shapes. Preliminary results allowed a subset of the estimated parameters to be identified that have shown good capabilities in discriminating the voice shapes. These parameters are related to distributions of voicing and silence periods, pitch and Cepstral Peak Prominence Smoothed. A web campaign has been also launched asking untrained subjects to “*give their voice to the research*”. Even though only two voice shapes have been identified in this data set, a comparison with the parameters extracted from the trained subjects has shown a good agreement.

Keywords: Voice analysis, Classification, Relational Communication

1. INTRODUCTION

The characteristics of the vocal signal has been used as important markers in different fields, as reported in the scientific literature. The speech quality in telephone systems has been evaluated analyzing the received voice signals (1,2). The assessment of voice disorders is often based on the value of parameters extracted from sustained vowels or continuous speech recorded using both microphone in air and contact microphones (3-5). The voice signal is also investigated in order to identify illness that are not related to the phonatory system, such as Parkinson diseases (6), depression and suicide risks (7) or obstructive sleep apnea (8).

Vocal-signal analysis has been also implemented in the classification of emotions (9-14). One of the first attempt in correlating voice parameters to anger, fear and sorrow is reported in (9), where different time and spectral characteristics of the speech signal have been estimated and a good correlation between emotions and contour of fundamental frequency vs time has been observed. Many works highlight that the identification of emotions from voice parameters is language dependent, as reported in (10) where a comparison between Mandarin and English is presented. Another example is reported in (11), where objective parameters have been evaluated for sentences characterized by different emotions in German and French and acoustic parameters have been extracted from the vocal signal using the software PRAAT. In this case, a multiple discriminant analysis has shown a good

¹ alessio.carullo@polito.it

² a.anibaldi@interagoacademy.it

³ arianna.astolfi@polito.it

correlation between objective and subjective evaluation of emotions. In (13) acoustic parameters related to fundamental frequency and perturbation quotient of amplitude and pitch period are extracted from vocal signals acquired during psychotherapy sessions in order to highlight anger and sadness. Another work (14) refers to the classification of ten different emotions in songs interpreted by eight opera singers in three different languages (English, French, German). The parameters defined in the GeMAPS (Geneva Minimalistic Acoustic Parameter Set) (15) have been extracted from the vocal signal and a multivariate analysis of variance (MANOVA) have shown a good discrimination between sadness and tenderness on the one hand and joy and pride on the other.

In this paper, a new relational paradigm is proposed that is based on the speech style that is used to establish an effective communication. The main goal of the research consists in objectively verifying this paradigm that associates a voice shape to each speech style, based on an original proposal of Interago Academy s.r.l., which is a company that works in the field of relational communication. To this aim, an attempt has been made to match the speech styles to acoustic parameters of the vocal signal, starting from data collected by 12 Italian actors, which have been trained to interpret five different voice shapes. Three functional voice shapes are defined: “rounded”, which means that a colloquial and empath communication is established; “triangular”, which means transmitting energy, joy and interest; “squared”, which highlights competence and solidity. Two dysfunctional shapes are also defined: “flat”, which is a monotone style that does not capture the listener attention; “spiky” or “X”, which is an aggressive style that transmits anger or blame towards the listener. The acquired vocal signals have been processed in order to extract several voice features, such as the histograms of occurrences of voicing and silence periods, pitch, intensity, and Cepstral Peak Prominence Smoothed (CPPS). This preliminary step allowed a set of parameters correlated to the voice shapes to be identified that act as references. Then, another data set has been processed that refers to 148 recordings of Italian untrained subjects, who have been asked to tell personal histories that evoke anger, happiness or fear. The parameters extracted from this second data set have been compared to the value obtained in the first step in order to evaluate the communication effectiveness.

In the following sections, the method for data collecting is described and the extracted acoustic parameters are defined. Then, a correlation-based technique is used to identify the parameters that are highly correlated to the voice shapes provided by the actors and a classification model is tested. Eventually, the results related to the untrained subjects are described and final comments are provided.

2. METHOD AND MATERIAL

2.1 Voice shapes

Five voice shapes are defined, which are subdivided in two classes. The functional class, which refers to speech styles that are effective from a relational point of view, includes the following shapes:

- ✓ rounded (**R**), which is used when the communication is characterized by colloquial and empath speech, e.g. when the goal of the communication consists in transferring emotions or private feelings to the listener;
- ✓ triangular (**T**), which highlights a high degree of involvement in the communication and the main goal is the transmission of energy, joy and interest;
- ✓ squared (**S**), which corresponds to a style that is used by people that show a high level of competence and solidity, as a well-prepared teacher or the speaker of a nature show.

The dysfunctional class includes two shapes that are not effective from a relational point of view:

- ✓ flat (**F**), which corresponds to a monotone style that is characterized by small changes in intensity, pitch and speech rate, thus resulting boring and not effective in capturing the listener attention;
- ✓ spiky (**X**), which is an aggressive style that transmits anger or blame towards the listener and does not match relational standards.

2.2 Subjects

The subjects involved in the experiment have been recruited in two different steps.

During the first step, 12 actors (7 males and 5 females) have attended a one-day training course based on a review of relational communication and the definition of the voice shapes. Then, each one has acted five passages (each passage about three-minute long) reproducing the five speech styles and

recording each passage with a smartphone in a quiet environment. These passages have been validated by the authors before considering them as “gold standards” for the five investigated voice shapes. During this validation phase, some of the actors have repeated the reproduction of some speech styles until the quality has been considered high enough from a subjective point of view.

In the second step, a web campaign has been launched asking voluntary subjects to “*give their voice to the research*”. The subjects involved in this second step have been asked to tell personal histories that evoke anger, happiness or fear and recording a passage of about three minutes using a web-based application accessible through smartphone, tablet or personal computer. Initially, about 400 contributions have been received and evaluated from a qualitative point of view, in order to discard recordings not long enough (less than one minute) and characterized by poor acoustic quality (high noise floor or presence of other sound sources). After this preliminary check, 148 contributions have been selected (19 males and 129 females), which have been processed according to the procedure described in the next section.

One should note that, due to the proposed recruitment procedure, information related to the quality of the recording system, such as bandwidth, noise floor and sensitivity, and the acoustics of the place where the recordings have been performed is not available. However, this does not represent a limitation for the reliability of the obtained results, since the final goal of the research is the development of a web-app that allows each subject to evaluate his/her voice using the microphone that is embedded in a smartphone or a tablet. The only requirements are a quiet environment and the absence of other acoustic sources. The main effect of the lack of acoustic requirements during the recruitment procedure consist in making worse the parameter reproducibility, which will be estimated and taken into account in the identification of the models for the voice-shape discrimination.

2.3 Acoustic parameters

The available vocal tracks have been processed in order to extract the parameters of interest. The same processing algorithms have been used both for the trained actors and for the untrained subjects.

A pre-processing has been performed by down-sampling the vocal signals to 22050 Sa/s, then the available samples have been grouped in 1024-point frames, which corresponds to a frame length of about 46 ms, and the root mean square value (RMS_{frame}), the harmonic-to-noise ratio (HNR_{frame}) and the pitch ($F0_{\text{frame}}$) of each frame have been estimated. Eventually, the analyzed frames are classified as “*voiced*” if the following conditions are met:

$$\left[RMS_{\text{frame}} > \frac{RMS_{\text{aver}}}{2} \right] \text{ AND } [HNR_{\text{frame}} > 0 \text{ dB}] \text{ AND } \left[\frac{|F0_{\text{frame}} - F0_{\text{frame-1}}|}{F0_{\text{frame-1}}} < 0.5 \right] \quad (1)$$

where RMS_{aver} is the mean value of the parameter RMS_{frame} along the whole recording. If one of the conditions expressed by the previous rule is not met, the frame is classified as “*unvoiced*”.

The voiced/unvoiced classification procedure allows a first set of parameters to be estimated, which are the percentage phonation time ($Dt\%$) and the histogram of occurrences of voicing and silence periods. The parameter $Dt\%$ is obtained as:

$$Dt\% = 100 \cdot \frac{n_{\text{voiced}}}{n_{\text{voiced}} + n_{\text{unvoiced}}} \quad (2)$$

where n_{voiced} and n_{unvoiced} are the number of frames classified as voiced and unvoiced, respectively.

In order to build the histograms of voicing and silence periods, the occurrences of uninterrupted sequences of voiced and unvoiced frames are counted. An example of histograms of occurrences of voicing and silence periods are shown in Figure 1, which refers to one of the untrained subjects. The information included in the histograms of voicing and silence periods is summarized by means of the parameters mode ($voiced_{\text{mode}}$ and $silence_{\text{mode}}$), mean ($voiced_{\text{mean}}$ and $silence_{\text{mean}}$) and standard deviation ($voiced_{\text{std}}$ and $silence_{\text{std}}$).

Only for the voiced frames, the histograms of occurrences of pitch and rms value are also estimated and the parameters mean ($F0_{\text{mean}}$ and RMS_{mean}) and standard deviation ($F0_{\text{std}}$ and RMS_{std}) are used to describe the behavior of each subject.

The voiced frames are also processed to estimate parameters in the cepstrum domain, that is, a log power spectrum of a log power spectrum. The first power spectrum shows how the signal energy is distributed among the frequency components, while the second one highlights the regularity of the harmonic components in the first spectrum.

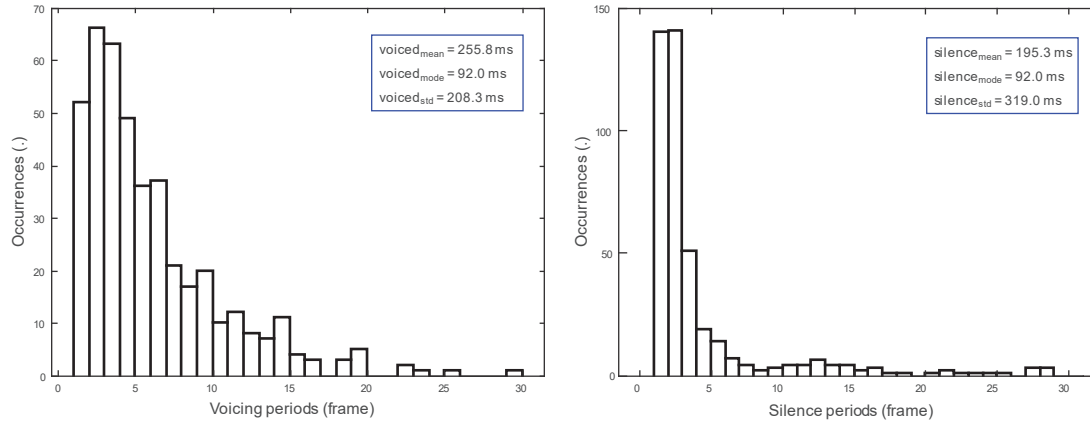


Figure 1 – Example of histograms of occurrences of voicing and silence periods for multiple duration of 46 ms, which is the frame length.

The parameter Cepstral Peak Prominence Smoothed (CPPS) has been estimated through a MATLAB script that the authors have developed according to the procedure described in (16). For each 1024-point (46 ms) frame, the Fast Fourier Transform algorithm is implemented twice to obtain the cepstrum. Then, a two-step smoothing procedure is performed: in the first step, a time smoothing cepstra is obtained along a time-window of 14 ms (seven shifts of 2 ms); in the second step, the smoothing of the cepstra magnitude is performed in the cepstral domain across quefrency using a seven-bin window. Once the smoothed cepstrum is obtained, a regression line is estimated in the quefrency versus cepstral magnitude domain excluding the first millisecond (17). Eventually, the parameter CPPS is obtained as the difference in decibel (dB) between the peak in the cepstrum and the value of the regression line at the peak quefrency. For each vocal track a CPPS histogram is obtained that includes a number of occurrences that is equal to the total duration of the voiced frames divided by 2 ms. As an example, if the voiced frames have a total duration of 60 s, the CPPS histogram will include $60000 \text{ ms} / 2 \text{ ms} = 30000$ values. An example of CPPS histogram is shown in Figure 2.

For each CPPS distribution, nine descriptive statistics are used: mean ($CPPS_{mean}$), median ($CPPS_{median}$), mode ($CPPS_{mode}$), fifth percentile ($CPPS_{5prc}$), and 95th percentile ($CPPS_{95prc}$), which measure the location of the distribution; standard deviation ($CPPS_{std}$) and the interval between the maximum and the minimum value ($CPPS_{range}$) as measures of the dispersion; kurtosis ($CPPS_{kurt}$) and skewness ($CPPS_{skew}$) that are related to the distribution shape.

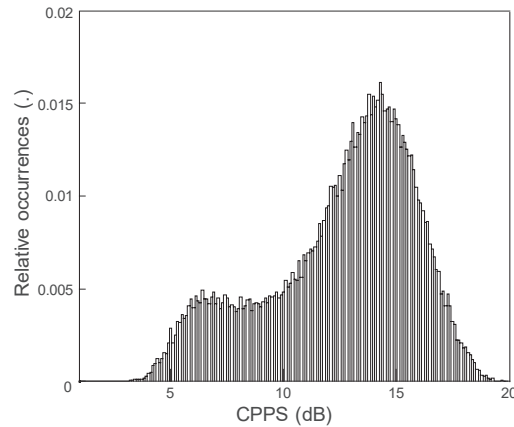


Figure 2 – Example of CPPS distribution.

3. VOICE-SHAPE CLASSIFICATION

3.1 Trained actors

The acoustic parameters extracted from the recordings of the trained actors have been associated to the gender of each actor and to the speech style that has been acted, obtaining a total of 22 features for each track: gender (1 feature), CPPS (9 features), voicing and silence periods (6 features), pitch (2 features), intensity (2 features), percentage phonation time (1 feature), voice shape (1 feature).

The correlation among the available features for the 60 available tracks (12 actors, 5 voice shapes) have been estimated through the Pearson's coefficient. Table 1 summarizes the estimated correlation coefficients between the voice shape and the other features: only the values reported as bold underlined are significantly different from zero ($p\text{-value} < 0.05$). It is then possible to state that the voice shape exhibits a medium correlation with the features $unvoiced_{std}$, $voiced_{mean}$, $unvoiced_{mean}$, $voiced_{std}$ and $F0_{mean}$, and a medium-low correlation with the features $F0_{std}$, $CPPS_{std}$, $CPPS_{5prc}$ and $Dt\%$. The other correlation coefficients cannot be considered reliable, since the size of the analyzed dataset is not large enough to confirm the null hypothesis "no correlation".

Based on this preliminary result, an attempt has been made to evaluate the voice-shape discrimination capability of the correlated features. Figure 3 shows, as an example, the features $voiced_{mean}$ and $F0_{std}$ of the 12 actors subdivided according to the different voice shapes using this coding: flat (magenta star), rounded (red circle), squared (blue square), triangular (green triangle), and spiky (black x-mark). The figure highlights the estimated correlation with respect to the voice shapes, which is negative for the feature $voiced_{mean}$ and positive for $F0_{std}$, even though the five classes are not well separated. A complete analysis of the features that are significantly correlated to the voice shape and that exhibit a correlation coefficient greater than 0.4 is summarized in the Figure 4, where each bar chart reports the mean value of the selected feature for the different voice shapes: flat (F), rounded (R), squared (S), triangular (T), and spiky (X). In the figure, the white bars refer to the results obtained for the trained actors, as also indicated by the vowel "a" added to each shape identifier. The error bar superimposed to each bar represents the interval (mean value \pm standard deviation of the mean). Some features seems to be good discriminators of voice shapes, such as $voiced_{mean}$ and $F0_{mean}$ since their intervals do not overlap to each other, while other features do not ensure a complete discrimination among voice shapes. In particular, the features $unvoiced_{std}$, $unvoiced_{mean}$ and $voiced_{std}$ exhibit similar values for flat and rounded shapes and the first two features do not allow triangular and spiky shapes to be distinguished. On the other hand, the feature $F0_{std}$ well distinguishes flat and round shapes, while the intervals for rounded and square shapes and for triangular and spiky shapes overlap each other.

The obtained outcomes suggest that a multiple-parameter model could be effective in the discrimination of the defined voice shapes. For this reason, the application "Classification Learner" that is embedded in the MATLAB™ environment (version R2018a 9.4.0.813654) has been used to identify a suitable classification model. Initially, simple decision-tree based algorithms have been tested using the six features previously selected and reported in the Figure 4. The best accuracy, which has been obtained without any validation scheme because of the small size of the data set, was 78.3% using a medium tree algorithm. In the same configuration, the accuracy reaches the value 80.0% if the feature $CPPS_{5prc}$ is added to the six selected features.

Table 1 – Pearson's correlation coefficient between the voice shape and the available features. Only the values reported as bold underlined are significantly different from zero ($p\text{-value} < 0.05$).

Gender	0.000	$CPPS_{95prc}$	-0.113	$unvoiced_{mode}$	-0.184
$CPPS_{mean}$	0.038	$CPPS_{skew}$	-0.080	$unvoiced_{std}$	<u>-0.679</u>
$CPPS_{median}$	0.037	$CPPS_{kurt}$	0.054	$F0_{mean}$	<u>0.598</u>
$CPPS_{mode}$	0.190	$voiced_{mean}$	<u>-0.635</u>	$F0_{std}$	<u>0.426</u>
$CPPS_{std}$	<u>-0.342</u>	$voiced_{mode}$	-0.202	RMS_{mean}	-0.067
$CPPS_{range}$	-0.027	$voiced_{std}$	<u>-0.598</u>	RMS_{std}	0.077
$CPPS_{5prc}$	<u>0.287</u>	$unvoiced_{mean}$	<u>-0.603</u>	$Dt\%$	<u>0.262</u>

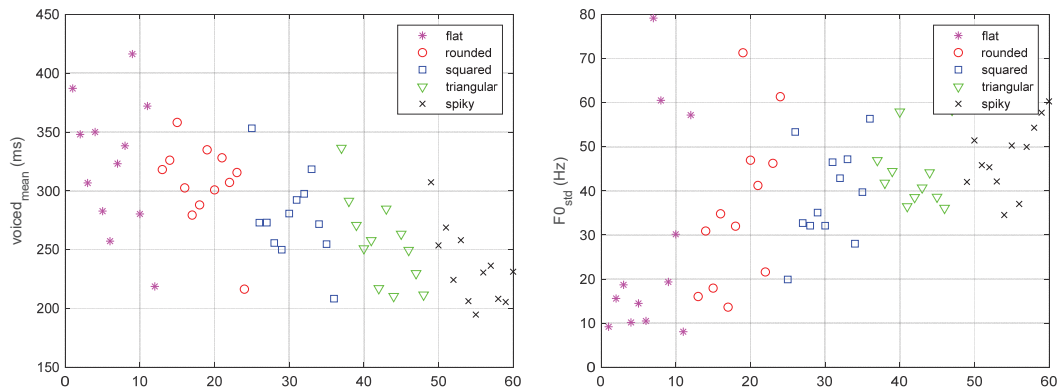


Figure 3 – The features $voiced_{mean}$ and $F0_{std}$ of the 12 actors subdivided according to the voice shapes.

The performance of this classification algorithm is summarized by means of the confusion matrix that is shown in the Figure 5, where the left chart shows predicted vs true voice shapes, while the right chart gives the corresponding true positive rate and false negative rate. One should note that the shapes with the best positive rate (92%) are squared and spiky, while the worst is the triangular shape with a positive rate of 67%.

Other kind of classification algorithms have been tested in the MATLAB environment using the six features that have been initially selected. With a K-Nearest Neighbor (KNN) estimator and with a Bagged Trees algorithm, a 100 % accuracy has been obtained in the classification of the voice shapes, however this result is considered unreliable due to the small size of the data set and the absence of any validation scheme.

3.2 Untrained subjects

The 148 tracks available for the untrained subjects (129 females and 19 males) have been initially evaluated from a subjective point of view in order to assign to each track the corresponding voice shape. Among the 148 subjects, 58 have told a history that evoke happiness, while 15 subjects have stated that their history refers to an angry situation. It is then expected that triangular and spiky shapes are present in the collected samples, but instead the subjective evaluation resulted in 76 flat shapes and 72 rounded shapes. This is a first important indication that highlights that untrained subjects are not able to establish an effective relational communication: the 58 “happy” subjects have produced 25 flat shapes and 33 rounded shapes, while the 15 “angry” subjects have produced 7 flat shapes and 8 rounded shapes.

After this subjective evaluation, the tracks of the 148 untrained subjects have been processed in order to extract the 20 features that have been described in the section 2.3. Then, the six features that have been selected during the analysis of the trained actors ($unvoiced_{std}$, $unvoiced_{mean}$, $voiced_{mean}$, $voiced_{std}$, $F0_{mean}$ and $F0_{std}$) have been processed estimating mean value and standard deviation related to the two identified voice shapes, which are flat and rounded. The obtained results are reported in the Figure 4 by means of gray bars, which are placed close to the same shapes of the trained actors. The most important outcome that arises from the comparison between the two categories of involved subjects is that the selected features are very similar for the voice shapes flat and rounded, and, with the exception of the feature $F0_{mean}$, the other five features have values that are very different from the squared, triangular and spiky shapes.

4. CONCLUSIONS

A new paradigm of effective communication has been defined that is based on the concept of “voice shape”. Five voice shapes have been proposed that are related to different communication patterns, distinguishing between functional (round, squared, triangular) and dysfunctional (flat, spiky) shapes. In this paper, the authors have made a first attempt in correlating the different voice shapes to acoustic parameters of the vocal signal. Starting from a sample of 12 actors that have reproduced the five voice shapes, the significantly correlated parameters have been identified.

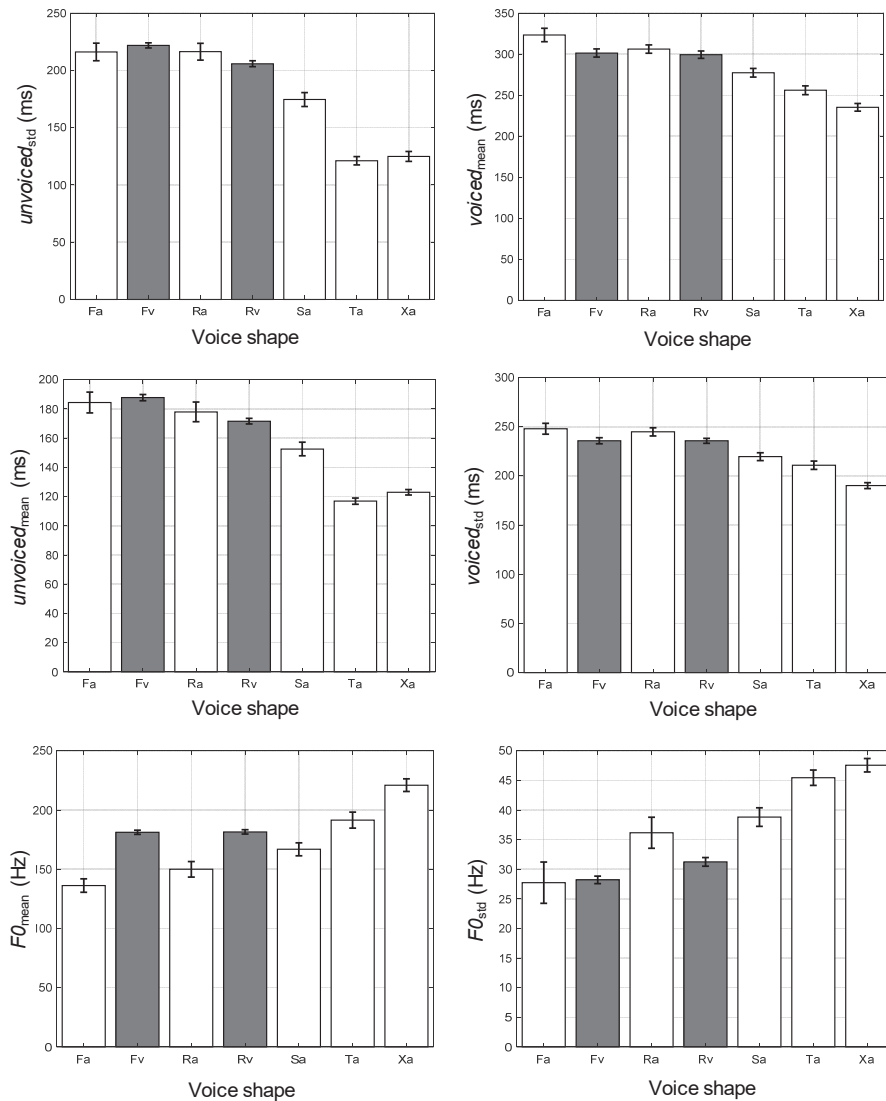


Figure 4 – Mean value and standard deviation of the selected features for the different voice shapes: flat (F), rounded (R), squared (S), triangular (T), and spiky (X). The white bars refer to the trained actors (“a” added to each shape identifier), while the gray bars refer to the voluntary untrained subjects (“v” added to each shape identifier).

A subset of the acoustic identified parameters has been then selected and used to train a medium-tree classification algorithm, which have provided an overall accuracy of 80%. However, this preliminary result suffers from the size of the available data set, which is not large enough to obtain meaningful outcomes. For this reason, the authors are enlarging the data set by recruiting other actors that will be trained to produce the different voice shapes.

Another experiment has involved 148 untrained subjects, who have been asked to tell personal histories that evoke anger, happiness or fear. However, these subjects were not effective from a relational point of view, since their voice shapes have been classified as flat or rounded, even when they told histories related to anger and happiness. Regardless of this poor relational capability, the acoustic parameters have been estimated also for the untrained subjects and compared to the same parameters of the trained actors. For the voice shapes flat and rounded, a good agreement has been obtained for five out of six parameters that have been previously selected.

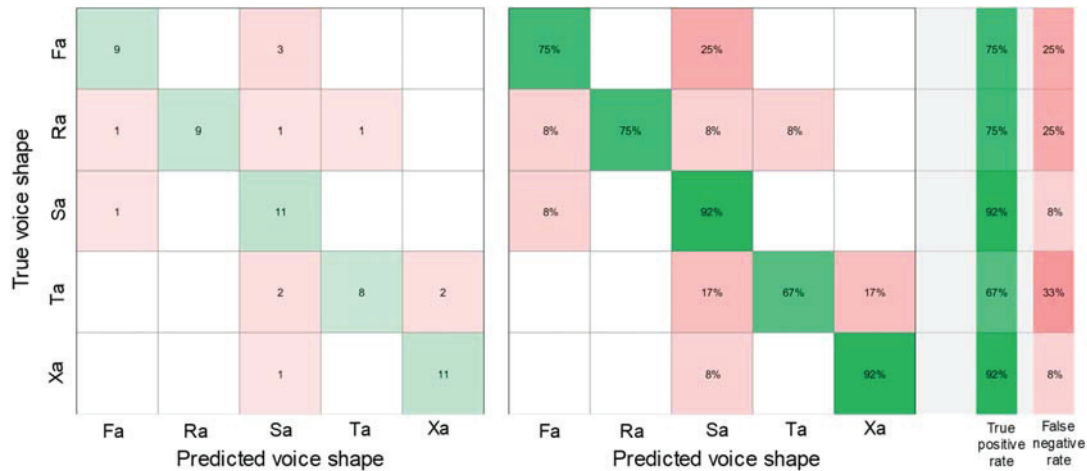


Figure 5 – Confusion matrix of a medium tree algorithm that is trained using the six selected features plus the feature $CPPS_{5pre}$: the left chart shows predicted vs true voice shapes; the right chart gives the corresponding true positive rate and false negative rate.

REFERENCES

1. Cai L., Tu R., Zhao J., Mao Y. Speech quality evaluation: A new application of digital watermarking,” IEEE Trans. Instrum. Meas. 2007;56(1): 45–55.
2. Carni D.L., Grimaldi D. Voice quality measurement in telecommunication networks by optimized multi-sine signals. Measurement 2008; 41(3): 266–273.
3. Parsa V., Jamieson D.G. Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. J. Speech Lang. Hear. Res. 2001;44(2): 327–339.
4. Maryn Y., Roy N., De Bodt M., Van Cauwenberge P., Corthals P. Acoustic measurement of overall voice quality: A meta-analysis. J. Acoust. Soc. Amer. 2009;126: 2619–2634.
5. Castellana A., Carullo A., Corbellini S., Astolfi A. Discriminating Pathological Voice From Healthy Voice Using Cepstral Peak Prominence Smoothed Distribution in Sustained Vowel. IEEE Trans. Instrum. Meas. 2018;67(3): 646–654.
6. Brabenec L., Mekyska J., Galaz Z., Rektorova I. Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation. J. Neural Transm. 2017; 124(3): 303–334.
7. Cummins N., Scherer S., Krajewski J., Schnieder S., Epps J., Quatieri T.F. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 2015;71: 10–49.
8. Herath D.L., Abeyratne U.R., Hukins C. Hidden Markov modelling of intra-snore episode behavior of acoustic characteristics of obstructive sleep apnea patients. Physiol. Meas. 2015;36(12): 2379–2404.
9. Williams C.E., Stevens K.N. Emotions and speech: Some Acoustical Correlates. J. Acoust. Soc. Am. 1972;52(4):1238–1250.
10. Wang T., Lee Y., Ma Q. Within and Across-Language Comparison of Vocal Emotions in Mandarin and English. Appl. Sci. 2018;8,2629:1–18.
11. Bänziger T., Patel S., Scherer K.R. The Role of Perceived Voice and Speech Characteristics in Vocal Emotion Communication. J. Nonverbal Behav. 2014;38:31–52.
12. Lee C.M., Narayanan S.S. Toward Detecting Emotions in Spoken Dialogs. IEEE Trans. on Speech and Audio Processing 2005;13(2): 293–303.
13. Rochman D., Amir O. Examining in-session expressions of emotions with speech/vocal acoustic measures: An introductory guide. Psychotherapy Research 2013;23(4): 381–393.
14. Scherer K.R., Sundberg J., Fantini B., Trznadel S., Eyben F. The expression of emotion in the singing voice: Acoustic patterns in vocal performance. J. Acoust. Soc. Am. 2017;142(4): 1805–1815.
15. Eyben F. et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans. on Affective Computing 2016;7(2): 190–202.
16. Hillenbrand J., Houde R.A. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. J. Speech Hearing 1996;39(2): 311–321.
17. Hillenbrand J., Cleveland R.A., Erickson R.L. Acoustic correlates of breathy vocal quality. J. Speech Hearing Res. 1994;37(4): 769–778.