Prosodic Parameters of Perceived Emotions in Vocal Server Voices

T. Ehrette, N. Chateau, C. d'Alessandro, V. Maffiolo

France Telecom R&D, 2, avenue Pierre Marzin – F-22307 Lannion Cedex (France)
LIMSI / CNRS, BP 133, Université Paris-XI – F-91403 Orsay (France)
{thibaut.ehrette; noel.chateau; valerie.maffiolo}@rd.francetelecom.com; cda@limsi.fr

Abstract

This paper investigates the relationship between perceptive and acoustic spaces describing a group of voices. The aim is to study possible correlations between prosodic parameters and the perceived quality of voices. Experiments are based on a set of 100 sentences recorded by 20 professional female speakers. A perceptual experiment aims at defining a set of relevant adjectives (or "attributes") for voice quality description (20 attributes are retained). Acoustic prosodic analysis is performed on the same corpus (23 parameters are used; mainly based on pitch, durations and energy). Then correlations between acoustic parameters and perceived attributes are computed. The results show that for some attributes, one can find a strong correlation with prosodic acoustic parameters (for about 13 attributes out of 20). However, for some attributes, these correlations are rather low, and more detailed investigations seem necessary in order to characterize perceived attributes in terms of acoustic parameters.

Keywords: voice identity, voice quality, prosodic parameters, telecom voices, voice quality perception.

1. Introduction

In telecom applications, voice is not only a way of giving semantic or linguistic information but it carries also non-linguistic information [5] such as the speaker identity, attitude or emotion. Due to the huge quantity of phone calls towards vocal servers and the great number of companies using servers as a gate for their customers, it seems important for telecom companies to control voice identity in order to better monitor their image and their marketing values.

The purpose of this article is to make a link between prosodic acoustic parameters and perceptive attributes of voice. As we used a same sentence, i.e. the same segmental content, for all the utterances in our corpus, it seems that emotion and voice identity is conveyed mostly by prosodic changes. It is then important to study if prosodic parameters, computed using signal processing, would correlate well with the results of perceptive judgments. But it is not clear how many acoustic parameters are needed to accurately characterize perceptive attributes, either alone or in combination with other parameters.

In a previous study [10], a set of adjectives, or "attributes", has been identified, according to a free categorisation and verbalisation test. These 20 attributes have been derived from the judgements of a group of naïve subjects whose task was to explain, using in their own words, their appreciation of vocal servers voices. Acoustic parameters have been chosen according to the results found in the literature dealing with acoustic correlates of voice emotion. Din spite of the diversity of experiments reported and of emotions studied, one can still find some regularity in the citations of factors and their range of values. This has guided our own choice of acoustic parameters.

Of course, F_0 seems to be the most relevant acoustic parameter. According to Williams [2] average values and ranges of F_0 are closely correlated with the emotional state of a speaker, in particular anger, sorrow and fear. However, according to Liebermann [1], in a study of eight different emotional modes (boredom, confidence, doubt, fear, happiness, objective question, objective statement and pompous statement), F_0 alone is not able to convey emotional information: at least signal amplitude has to be added. However, in a large multi-speaker corpus, a lot of care has to be taken because of the large differences in the range of F_0 and amplitude among different speakers [4]. Moreover, different speakers are using different strategies, and different parameters to express the same emotion. Hirose [7] shows that, even if there are variations in the manner to express attitude or emotion, prosody remains the main indicator. Few studies consider also temporal aspects and are often limited to the comparison of global duration time variation through the different emotions. Mozziconacci [9] examines the mean global speech rate and points out the importance of this parameter for characterizing emotions like boredom.

The paper is organized as follows. In section 2, the speech corpus is presented, the 20 perceptive attributes derived from these data are described, and the 23 acoustic measures used are explained. Section 3 presents the statistical analysis of correlation between perceptive attributes and acoustic parameters. Section 4 discusses of the results and concludes.

2. Corpus, perceptual and acoustic analyses

The corpus contains 100 speech utterances recorded in a professional studio, all containing a same sentence: *"Bienvenue sur Audiotelis; pour obtenir dès maintenant le service de votre choix, tapez la commande correspondante sinon suivez-moi." ("Welcome on the Audiotelis server: to select the service you have chosen, press the corresponding key, or follow me.").* Twenty professional female speakers were asked to record the messages according to their own interpretation of five speaking styles: *normal ("naturel")*, *warm ("chaleureux")*, *dynamic ("dynamique")*, *reassuring ("rassurant")* and *smiling ("souriant")*.

2.1. Perceptive adjectives

A free categorization method, described in [6-8] and derived from [3], has been used for perceptual analysis. Five shares of the 100 voices corpus, composed by 40 sentences, are respectively presented to 10 groups of 17 to 20 subjects. A grand total of 185 male and female subjects ageing between 14 and 60 years old participated in the experiments. On a computer screen, each subject visualises 40 “balls” representing the 40 sounds. They can listen to the voices as often as they want, the sound restitution is made by telephone handset. Subjects are asked to group the voices, by dragging and dropping the balls with the mouse, according to their own feeling of similarity. There is no restriction on the number of groups or the number of items in each group. After grouping, the subjects are asked to describe freely each group, with their own words. The 20 most frequent words are derived from these descriptions: *“accueillante” (welcoming)*, *“agréable” (pleasant)*, *“agressive” (aggressive)*, *“autoritaire” (authoritarian)*, *“banale” (ordinary)*, *“chaleureuse” (warm)*, *“claire” (clear)*, *“criarde” (shrill)*, *“dynamique” (dynamic)*, *“exagérée” (exaggerated)*, *“expressive” (expressive)*, *“gaie” (happy)*, *“jeune” (young)*, *“naturelle” (natural)*, *“professionnelle” (professional)*, *“rapide” (speedy)*, *“rassurante” (reassuring)*, *“sensuelle” (sensual)*, *“souriante” (smiling)* and *“stressante” (stressful)*. There is certainly redundancy in these terms, for example a *dynamic* voice may be *speedy*, an *aggressive* voice may not be *pleasant*, but it is not embarrassing for statistical treatment. The result of this experiment is a set of attributes, or perceptual qualities, that are spontaneously associated with the voice qualities in the 100 sentences of the corpus. In a second test, subjects were asked to score each attribute, on a scale from 0 (not effective) to 6 (perfectly matched) for each voice of the corpus. These numeric values are then converted from 0/6 to -10/+10 and they will serve in Section 3 for statistical analyses.

2.2. Acoustic analyses

Acoustic analyses are performed using 23 parameters that are summarised in Table 1.

F_0	ENERGY	DURATION	QUALITY
Average	Average	Speech sequence duration	Colour
Minimum	Maximum	Total silences duration	Spectral centroïde
Maximum	Variance	Silence-speech rate	
Range	Skewness	Pauses number	
Variance	Kurtosis	Average silence duration	
Skewness		Maximum silence duration	
Kurtosis		Silence duration variance	
		Silence duration skewness	
		Silence duration kurtosis	

Table 1: The 23 prosodic and voice quality parameters.

These parameters are describing prosodic and voice quality aspects of speech: pitch, energy, duration, and long-term spectrum. F_0 is computed every twenty milliseconds F_0 values are from 148 Hz to 252 Hz. Energy is calculated every ten milliseconds. All pauses and micro-pauses are signal segments where the RMS-energy is below an arbitrary 0 dB threshold, chosen in order to interpret breathings as pauses and to detect the lowest possible speech segment. Instantaneous pitch and energy values are difficult to use as such, therefore statistical values are computed, such as the average, variance, skewness and kurtosis. Skewness represents the symmetry of the distribution and kurtosis its sharpness. A small set of measures is then describing each utterance. Rhythmic aspects are difficult to analyse and, at a first glance, only temporal aspects will manage to describe it: silence information (speech portions energy below 0dB) and total speech sequence duration. Two long-term spectral descriptors are used to describe voice quality: *Spectral centroïde* and *Colour*. *Colour* is defined as the logarithm of the ratio between geometric and arithmetic average of the power spectral density distribution of the signal. It is taken as an indicator of difference between a spectrum with sharp harmonic peaks and a more flat one.

3. Perceptual/acoustic correlation

3.1. Statistical analysis

Standard correlation between all acoustic parameters and each perceptual attribute are computed to analyse the relationship between acoustic parameters and voice perception. Based on one or two best-correlated parameters, the hundred voices are then mapped on a 2D acoustic-perceptive graph. Some regularity can be noticed: the majority of voices seem to follow the same types of rules. However, other voices show different behaviours, and should be more carefully examined, with the help of detailed listening. From this last practical experience, new acoustic parameters may be necessary.

3.2. Analysis of multiple regression

Then, multiple regression can be used to analyse the relationships between the acoustic measures and the perceptual attributes. For each attribute, a statistical model can be designed, using 1 to 23 parameters. These models should be able to predict each perceptual attribute.

Figures 2 and 3 represent the so-called “observed variables” (i.e. scores given by subjects for the “*dynamic*” attribute during the perceptive tests) and the so-called “predicted variables” (i.e. scores predicted by the model for the same attribute). Each point represents a voice. The more the voices are grouped on the diagonal line, the more the model can be considered good and the better is the prediction.

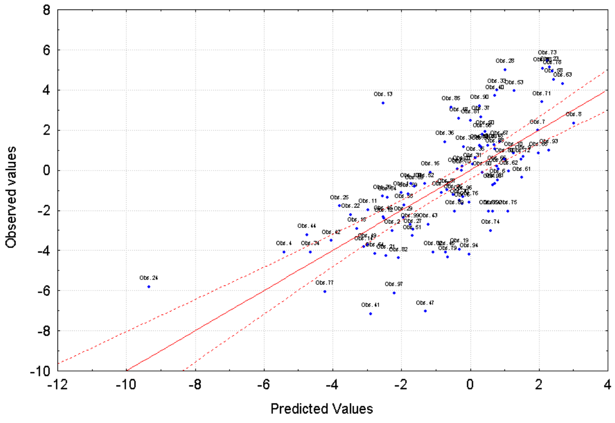

Figure 2: Prediction of the “dynamic level” by the “total duration time”, for the 100 voices.

In Figure 2 for instance, one can see the “*dynamic* level” of the voices predicted with the “*total silence duration*” parameter, which appeared as the best correlated parameter. For this example, only one parameter is able to explain

49,41% of the original variability (R²). With the first three best-correlated parameters, R² reaches 56,42%. When using all the acoustic parameters, a maximum R² of 75,60% is obtained. This is the limit of prediction of the "*dynamic"* attribute according to this set of acoustic parameters. However all these acoustic parameters are not very useful, as some of them do not affect strongly R². For this example *colour*, *spectral centroïde*, *energy kurtosis*, *silence duration variance* and *skewness*, and *minimum* F_0 can be removed from the model, keeping R² above 75%. Figure 3 is the same as Figure 2, but with a set of 17 parameters. The picture is clearly better, as all voices are better concentrated.

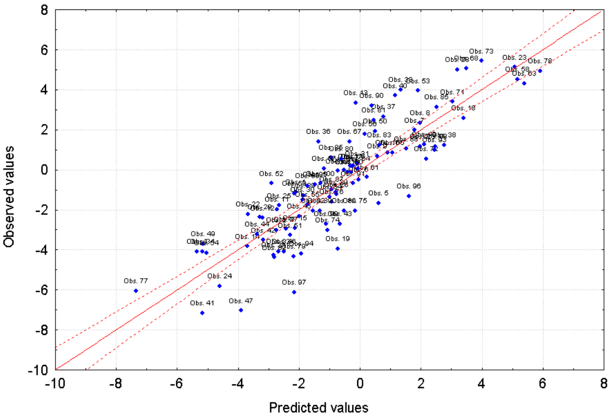

Figure 3: Prediction of the "dynamic level" by 17 acoustic parameters (R²=75,39%), for the 100 voices.

Table 4 shows all R² values for each perceptive descriptor including a total of 23 acoustic parameters (column 2) and only with the best-correlated parameter (column 3). The last column indicates the number of parameters for R² remaining above 50%.

An arbitrary limit of 50 % is chosen as a reference for a "good prediction". Only 13 attributes are above this limit. For some of them (e.g. *speedy*, *shrill*, *dynamic*, *aggressive* and *stressful*), only few acoustic parameters are needed, as they can be correctly predicted using 3 parameters or less. On the contrary, for other attributes, more parameters are needed.

Table 4 highlights two types of parameters, related to duration and the long-term spectrum. It seems that pitch or amplitude parameters are only of secondary importance in this analysis. The opposite result has been obtained for perceptual and acoustic analyses of syllabic sized or very short sentences in [11], where sounds where mainly grouped according to pitch and energy. However, some attributes like, "*authoritarian*" can be well predicted (R²=50,09%) by a combination mainly composed of fundamental frequency parameters: *authoritarian* = -3,49 - 0,28 x F_0*_average* - 0,09 x F_0*_minimum* + 2,14.10^3 x F_0*_maximum* - 2,05.10^3 x F_0*_range* - 0,11 x F_0*_kurtosis* + 0,98 x *average_energy* - 1,60 x *speech_sequence_duration* + 0,97 x *spectral_centroïde* + 0,73 x *total_silences_duration*. It can be noticed that these parameters may depend closely on the particular sentence used in this experiment. For instance, "*total duration time"* is specific to the particular prompt used, and may be used only for comparison in the context of this prompt. Another problem with a global parameter like "*total silence duration"* is that it does not explain anything about the breakdown of pauses during the utterance. A same value could be obtained with only one long pause, or many short pauses. Even if it shows alone a high correlation score, this kind of parameter has no absolute meaning, depends on other parameters, and of course depends also on the sentence used. *"Spectral centroïde"* appears to be the best-correlated parameter for the majority of perceptual attributes (12/20). This parameter describes the global spectral balance of the voice, which shows a high positive correlation with "*shrill"*, "*aggressive"*, "*stressful"* and a high negative correlation with "*sensual"*, "*pleasant"*, "*reassuring"* and "*warm"*.

Perceptive adjectives	**R² (%) with**		**Nbr R²>50%**
	23 parameters <	**> the best parameter**	
Speedy	79.91	51,19 (total duration time)	1
Shrill	77.28	45,80 (spectral centroïde)	2
Dynamic	75.60	49,41 (total silence duration)	2
Aggressive	70.81	30,88 (spectral centroïde)	3
Authoritarian	66.68	25,10 (pauses number)	9
Stressful	65.26	45,27 (spectral centroïde)	2
Sensual	63.92	34,14 (spectral centroïde)	8
Pleasant	58.29	38,85 (spectral centroïde)	7
Expressive	57.81	35,14 (silence-speech rate)	9
Young	54.54	14,93 (average silence duration)	15
Reassuring	51.81	36,54 (spectral centroïde)	12
Warm	50.58	28,06 (spectral centroïde)	19
Happy	50.43	22,25 (silence-speech rate)	20
Natural	49.78	14,44 (spectral centroïde)	+
Exaggerated	48.56	15,49 (spectral centroïde)	+
Welcoming	48.32	21,04 (spectral centroïde)	+
Professional	45.68	13,12 (total silence duration)	+
Smiling	42.03	16,16 (silence-speech rate)	+
Ordinary	38.82	9,02 (spectral centroïde)	+
Clear	35.24	9,18 (spectral centroïde)	+

Table 4: R² for each perceptual attribute

3.3. Analysis of simple correlation

Multiple regression seems not sufficient to explain the strategy used by subjects to judge voice quality. More analysis is needed to determine which acoustic parameters are good indicators of voice quality, and if they can be used in the same way for all voices.

In the previous paragraph, we have seen that "*speedy"* could be accurately predicted with only one parameter, i.e. the total duration time of the utterance. It seems to be a good predictor for most of the voices: the shorter is the duration, the speedier is the voice. It is a simple and natural concept and it appears also clearly in the statistical analyses. However many voices follow a different rule, as can be seen on the following example (Figure 5).

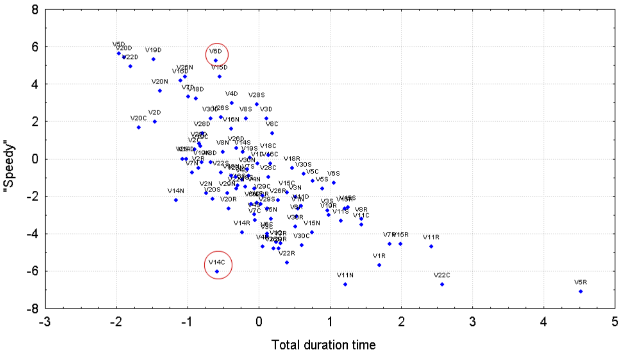

Figure 5: Correlation between "total duration time" and "speedy". VD6 and V14C are in a red circle.

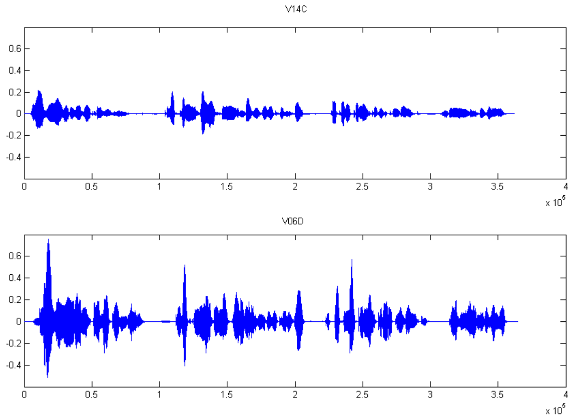

Figure 6: The two voices signal

Figure 6 shows two voices with the same duration time, that have a high score (V6D) or a low score (V14C) for the attribute "*speedy*". When observing the acoustic signal, one can notice the same number of pauses, at the same places and with the same length. When listening to both utterances it is clear that V6D is tenser while V14C contains smooth consonantal attacks, with no sharp articulation. This shows that some kind of "expert" listening is also needed, and that new acoustic descriptors have to be searched for, particularly in terms of articulation and voice source parameters.

4. Discussion and conclusion

Attributes like "*speedy*", "*shrill*", "*dynamic*" are closely connected to few acoustic parameters. There is certainly a masking effect on other parameters that do not affect seriously the perceptive judgment.

More vague attributes, like "*ordinary*" voice or "*professional*" voice, reach rather bad R^2 scores. It is difficult to understand the correspondence of such high level descriptions and acoustic parameters, at least with the current set of parameters. It should be necessary to take also into account more detailed information, e.g. pitch and formants on specific vowels. Along the same line, "*smiling*" is badly explained even though it is relatively well understood in terms of production (position and amplitude of the 3rd formant).

Therefore, it should be also necessary to use some segmental analysis, and a more refined spectral analysis on specific segments. Prosodic parameters may not be sufficient to characterise all aspects of perception of the speaking style. The originality of this work comes from the large number of acoustics parameters and the large number of voice examples as well as the statistical methods enabling to match acoustic and perceptive spaces.

In summary, we found that the 23 acoustic parameters can be well correlated with only a part of the 20 perceptual attributes (13 out of 20). They are not sufficient to explain more vague and subtle concepts that subjects spontaneously use, like "*ordinary*" or "*professional*". Some attribute may also be correlated more directly to production parameters, like e.g. "*smiling*", which do not correlate strongly with any prosodic aspect.

Two aspects seem important for future work. On the one hand, more acoustic parameters are needed. For instance voice source parameters (e.g. glottal open quotient, periodic-aperiodic ratio, spectral tilt) and articulation parameters (e.g. formants, speed of articulation). On the other hand it seems that expert listening is also needed, because free verbalisation of naive subjects may result in too vague and intricate attributes.

5. References

[1] Lieberman, P.; Michaels, S.B., 1962. Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech. *JASA*, 34, 7, 922-927.

[2] Williams, C.E.; Stevens, K.N., 1972. Emotions and speech: some acoustical correlates. *JASA*, 52, 4 part II, 1238-1250.

[3] Dubois, D., 1991. *Sémantique et Cognition - Catégories, prototypes, typicalités.* Paris: CNRS Editions.

[4] Stone, R.E.; Rainey, J.a.C.L., 1991. Intra- and Intersubject Variability in Acoustic Measures of Normal Voice. *Journal of Voice*, 5, 3, 189-196.

[5] Eckert, H.; Lavers, J., 1994. *Menschen und ihre Stimmen - Aspekte der vokalen Kommunikation*. Beltz.

[6] Guyot, F., 1996. *Etude de la perception sonore en termes de reconnaissance et d'appréciation qualitative: une approche par la categorisation*. PhD thesis, Le Mans Uiversity, France.

[7] Hirose, K.; Kawanami, H.; Ihara, N., 1997. Analysis of Intonation in Emotional Speech. *ESCA Workshop on Intonation: Theory, Models and Applications*, 185-188.

[8] Maffiolo, V., 1999. *De la caractérisation sémantique et acoustique de la qualité sonore de l'environnement urbain*. PhD thesis, Le Mans University, France.

[9] Mozziconacci, S.J.L.; Hermes, D.J., 2000. Variations temporelles communiquant l'émotion dans la parole. *XXIII journées d'étude sur la parole*, Eindhoven, Pays Bas.

[10] Maffiolo, V.; Chateau, N., 2001. Speech's Emotional Quality in Vocal Services. Proceedings of the *International Conference on Affective Human Factors Design*, Singapore.

[11] Payri, B., 1999. *Perception de la voix parlée: cohérence du timbre du locuteur*. PhD thesis, University Paris XI, France.