MODELLING THE EMOTIONAL QUALITY OF SPEECH IN A TELECOMMUNICATION CONTEXT

Noël Chateau *Valérie Maffiolo* *Thibaut Ehrette* *Christophe d'Alessandro*

Human Interactions Direction, France Telecom R&D
2, avenue Pierre-Marzin, 22307 Lannion Cedex, France

noel.chateau@francetelecom.com valerie.maffiolo@francetelecom.com thibaut.ehrette@francetelecom.com

Perception située
LIMSI / Cnrs, BP 133
Université Paris-XI
91403 Orsay, France
cda@limsi.fr

ABSTRACT

This paper presents a study of the perception, the analysis, and the modelling of the emotional quality of speech. Speech emotional quality is defined as the qualities of speech samples in terms of the emotional content that describe the listeners' global impressions as elicited by their audition. For this study, twenty professional female speakers recorded a welcome prompt of a vocal server in five elocution styles. The sound corpus was submitted to psychoacoustic tests and to signal analysis. From the psychoacoustic tests, twenty subjective criteria could be extracted that characterize the perceived emotional quality. These criteria can be used to draw perceptive portraits of the speech samples. Linear models connecting the perceptive portraits to physical data derived from signal analysis were developed.

1. INTRODUCTION

Due to the increasing applicability of speech-based technologies (speech recognition, speech synthesis, management of large pre-recorded speech databases, VoiceXML platforms…), telephone vocal servers are becoming a predominant way of accessing a large range of information, from voicemails to weather forecasts, and of course, the world wide web with the development of voice browsers (http://www.w3.org/Voice/). More than ever, this development of vocal servers reveals that speech material is the main medium between telecommunication operators and their customers.

The creation of a vocal server for a telecommunication operator follows three basic rules. First, on the conceptual level, the server must be useful for the targeted customers and must offer relevant information. Second, on the ergonomics level, the server needs to be user-friendly, requiring only a minimum of practice. Far too many servers are described by users as "a labyrinth of invisible and tedious hierarchies" [1]. Third, on the "look & feel" level, the server should convey a particular style, expressing the brand's value, and should be adapted to the targeted customers. If the first two rules are generally taken into account by telecommunication operators, the third is often forgotten or left to sound-recording studios that propose trendy speakers and music to create a particular audio environment. Unfortunately, the separation between the conceptual and ergonomic design on the one hand, and the "look & feel" design on the other is not very logical. The pragmatic content expressed by the style can be of great help in understanding the semantic content of speech signals by resolving ambiguities [2]. Moreover, it might be dangerous to rely upon the trendy attitudes of studios in selecting professional speakers and music without taking into account customers' perception and tastes in the design process.

This paper presents a study of the perception, the analysis and the modelling of styles or the "emotional quality" [3] of speech samples. Its aim is to construct a tool for vocal server designers that would help them to manage the vocal design of their services. First, we present the sound corpus that was studied, how it was recorded, submitted to psychoacoustic tests and to signal analysis. Then, a first model connecting subjective to physical data is proposed. This model was implemented in a web-based tool easy to use by designers, which will be demonstrated during the ICAD conference.

2. SOUND CORPUS

The corpus was constituted of one hundred vocal sequences and was established by asking twenty professional female speakers to pronounce one prompt for a vocal server named Audiotelis. This prompt was: "Bienvenue sur Audiotelis. Pour obtenir dès maintenant le service de votre choix, tapez la commande correspondante, sinon, suivez-moi." It was recorded in five elocution styles. These styles were chosen as representative of those of interest in a vocal server: natural, warm, dynamic, reassuring, and smiling. In all cases, speakers were asked to record the messages according to their own interpretation of what a natural, warm, dynamic, reassuring or smiling style was.

3. PSYCHOACOUSTIC TESTS

3.1. Free Categorization and Verbalization Test

Two series of tests were conducted on the corpus. The first series consisted of a free categorization and a free verbalization qualification of the sound corpus by one hundred and eighty-five listeners. This method, initially developed by Dubois [4] in the visual and tactile fields, was used so as not to impose any *a priori* constraints on the listeners that would have allowed them to identify the parameters of construction of the sound corpus, thus guiding their perceptive activities as responses to these parameters. As the sound corpus was very large, it was split into five groups of forty speech samples, each group being assessed by seventeen to twenty listeners. During the test, the listeners could listen to the speech samples represented by balls on a computer screen by clicking on them. Their task was first to group the speech samples (the balls) into coherent groups in terms of the general impressions they had when they heard them, and then to verbally qualify the groups.

For each of the five tests, a distance matrix could be computed between all pairs of speech samples. The distance between two samples is directly proportional to the number of times these two samples were grouped together by subjects. This distance matrix was analyzed by a tree algorithm [5]. Vertexes and vertices that join vertexes to each other characterize a tree. A vertex of order one is called a leaf, and a vertex of order >1 is called a node (the degree of a vertex being equal to the number of vertices connected to it). In this tree representation, leaves correspond to the vocal sequences. Nodes reflect the groupings. The lengths of the vertices reflect dissimilarity relationships between stimuli: if the length between two stimuli is short (resp. long), this means that these two stimuli are perceptually close to (resp. far from) each other. The number of nodes between two stimuli also reflects the degree of similarity between them: the higher the number of nodes, the less similar the stimuli. Figure 1 gives an example of a tree obtained from one of the five tests. The number of each "leaf" is the number of the speaker in the database and the letter refers to the elocution style with the convention: /n/ for natural, /d/ for dynamic, /s/ for smiling, /r/ for reassuring and /c/ for warm. It can be seen in Figure 1 that some speech samples of the same speaker are grouped together (*e.g.* Speaker 28), indicating that listeners perceived only small differences in the different elocution styles she produced. On the contrary, some speakers have their five recordings totally split and far apart from each other (*e.g.* Speaker 22), indicating large differences from one style to another.

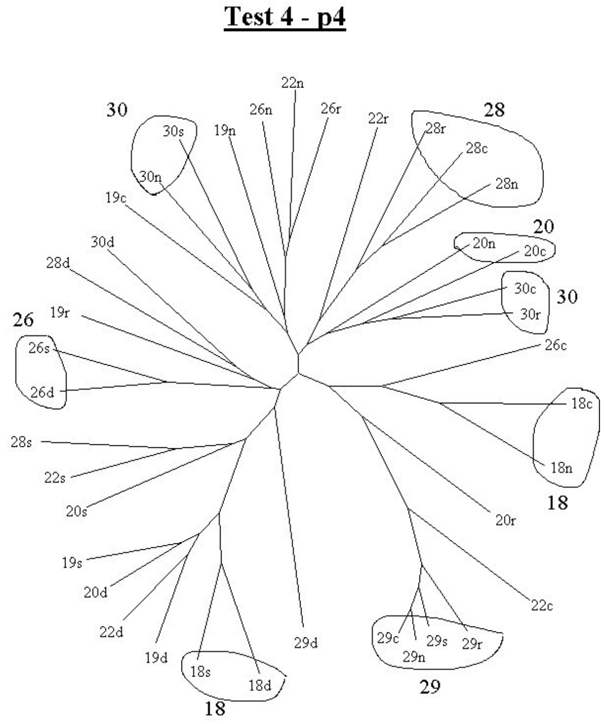

Figure 1. *A tree representation of the corpus for one test*

The global analysis of the five trees shows that some speakers are grouped, some semi-grouped and others split. This representation allows for the identification of coherent speakers whose elocution styles vary only slightly and who do not produce stereotypes of the desired styles that might be perceived as caricatures.

A semantic and frequency analysis of the verbal comments of the listeners was conducted. Interestingly, these verbal comments concern the speaker herself (*e.g.* "the person is happy to welcome us"), the speaker through her voice (cheerful, vigorous, sympathetic), the voice of the speaker with phonetics/acoustics terms (*e.g.* "clear, good intonation and good rhythm"), the intention of the speaker (*e.g.* "considerate, willing to please"), and the effect on the listener (*e.g.* "I understand there is an error but I do not feel guilty about it", "It gives the impression that…"). This is certainly the consequence of our instructions, which asked subjects to comment on their impressions, without guiding them more precisely.

In order to compare the voices in a common reference frame, it is interesting to construct a multidimensional space where they can be organised. To obtain this space, a frequency analysis of the occurrences of subjects' wordings was conducted. Twenty criteria could be derived from this analysis. These criteria were used in the second experiment to analytically assess the perceived emotional quality of our corpus.

3.2. Multi-Criteria Bipolar-Scales Test

In this test, fifty-four listeners had to assess the whole corpus according to the twenty criteria extracted from the first test. These criteria were presented in the form of 7-points scales from "pas du tout" (not at all) to "extrêmement" (extremely). Figure 2 shows one of the twenty scales used in the test, for the "agréable" criterion (pleasant). Each sequence was presented four times, one after the other. During the presentation, subjects had to score each scale even if some of them seemed to be less relevant.

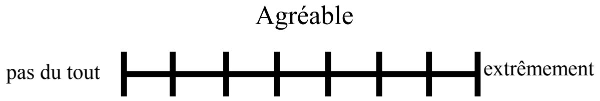

Figure 2. *An example of a bipolar scale.*

This method is often used to study sound perception. Its aim is to evaluate the contribution of different semantic attributes on perception. It allows the experimenter to easily direct the subjects' activity. It is simple for the subjects and doesn't require the construction of bipolar scales, thus eliminating the problem of choosing contrasting adjectives belonging to the same dimension. [6].

For each speech sequence and for each criterion, the numeric values obtained were averaged over the fifty-four subjects. This averaging results in a perceptive portrait of each vocal sequence. Figure 2 gives an example of the perceptive portraits for two speech samples produced by the same female speaker, one in the "warm" style (diamonds) and the other in the "dynamic" style (squares). For easier use, numerical values of the scales ranging from 1 to 7 were converted to a scale ranging from –10 to +10. On this kind of spider web, the center means "not effective" and the periphery means "perfectly matched". First of all, it can be seen that the warm and the dynamic criteria obtain values that correspond with the interpretation of the speaker. However, other criteria are very sensitive to style, and it might be observed that a dynamic style is also perceived as more aggressive and authoritative, more exaggerated, rapid, shrill, and stressful. On the contrary, a warm style is perceived as more pleasant, reassuring, and professional. Some criteria obtain similar values for both styles: smiling, young, expressive, cheery, and clear. These criteria are rather independent of the style (for the two styles represented here) and may refer to intrinsic qualities of the speaker (for example, whatever the style, a voice may appear more or less young or clear).

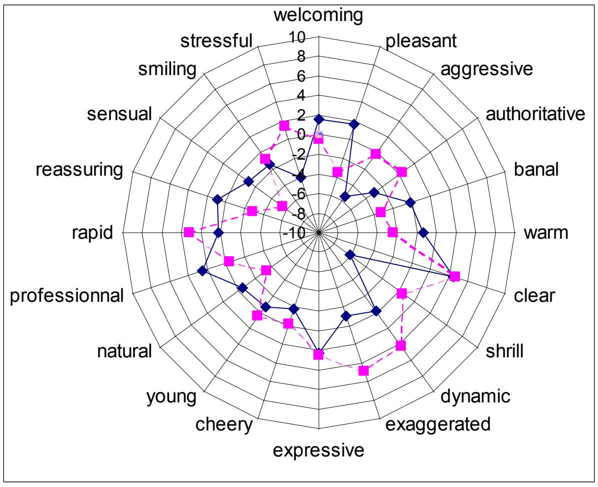

Figure 2. *Perceptive portraits for two speech samples for the same speaker and for two different elocution styles (warm = diamonds and dynamic = squares)*

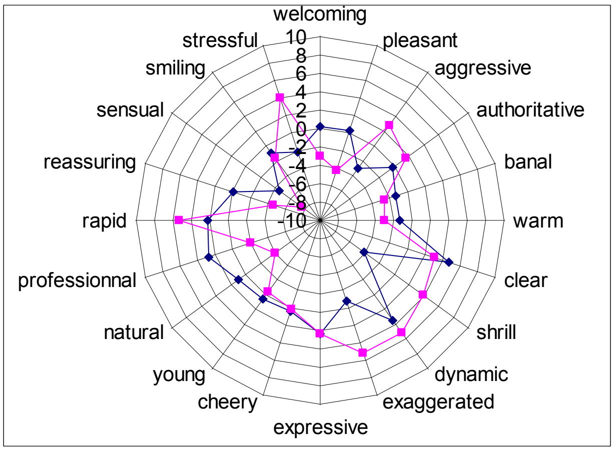

Figure 3. *Perceptive portraits for two speech samples for the same style (dynamic) and for two different speakers (Speaker 3 = squares, and Speaker 6 = diamonds).*

Figure 3 gives another example of two perceptive portraits of two different speakers (Speaker 3 = squares and Speaker 6 = diamonds) for the same elocution style (dynamic). It can be observed that although the two speakers are rather close for the "dynamic" criterion, Speaker 3 appears to be more unpleasant than Speaker 6, more authoritative and aggressive, much more shrill, more exaggerated, more rapid, and much more stressful. Actually, it seems that Speaker 3 exaggerated her interpretation of the dynamic style, which was consequently perceived in a negative way. On the contrary, it appears that Speaker 6 could express a dynamic style without being particularly stressful and unpleasant.

These two graphics show the large spread of the subjective data that is the consequence of intrinsic and extrinsic differences among speakers.

On the application side, a graphic search engine was devised in order to find speech samples in the database according to subjective criteria. The user is invited to draw the perceptive portrait of the target voice for the service she/he is looking for; the algorithm then finds the five speech samples in the database that best match the target. He/she can compare the different perceptive portraits, listen to the samples, and get more information about the recordings, such as data on the professional speaker.

3.3. Principal Component Analysis of the Data Set

Although the twenty criteria are the most representative of the verbal comments given by subjects in the first test, it can be hypothesized that this set of criteria can be reduced, since some of them refer to similar concepts (*e.g.*, aggressive, authoritative and stressful, dynamic and rapid). The correlation coefficients between the twenty criteria were computed over the one hundred speech samples. It can subsequently be concluded that the following subsets of criteria are highly correlated ($r > 0.9$):

- welcoming/pleasant/warm/reassuring
- aggressive/authoritative
- smiling/expressive/cheery
- clear/professional
- dynamic/rapid
- stressful/shrill
- warm/sensual

In addition, natural/exaggerated are correlated below –0.9. On the contrary, the "young" and "banal" criteria are correlated to all other criteria below 0.2, which indicates a specificity common to these two criteria. As many criteria were highly correlated, it became of interest to conduct a main component analysis in order to try to reduce the number of relevant criteria that significantly contribute to the identity portrait of each speech sample (see [7], for example, for a detailed description of the principal component analysis). Such an analysis was conducted. Below, Figure 4 shows the scree plot of the eigen values extracted from the analysis. Each eigen value is associated to a principal component (PC) that is a vector of the reduced basis where the data can be represented.

The sum of the values of the eigen values is equal to the global variance of the data set. The contribution of each eigen value (and therefore the corresponding PC) to the explanation of the variance of the data set can simply be computed as the ratio of the value of the eigen value to the sum of all eigen values. The resulting percentages appear in Figure 4 where it can be seen that the first main component accounts for 43.71% of the variance of the data set, the second for 27.57%, the third for 16.25%, and finally the fourth for 3.54%.

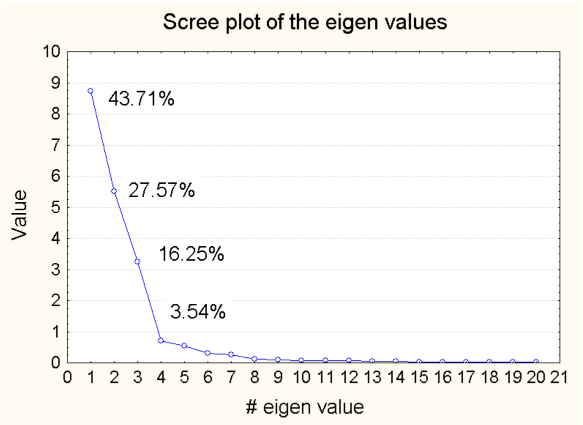

Figure 4. *Scree plot of the eigen values obtained from the principal component analysis of the data set.*

The classic interpretation of a scree plot is to select those eigen values which appear before the first "elbow" of the plot. These eigen values significantly contribute to the explanation of the global variance of the data set, whereas those located after the elbow can be considered as noise. It can be seen in Figure 4 that four eigen values should be retained; they account for 91% of the variance of the data set, which is a very good result.

Figure 5 and 6 respectively show a projection of the twenty criteria in the new basis formed by the PCs in the plans PC1 x PC2 and PC3 x PC4. The coordinate of each criterion on each PC is simply the correlation coefficient between each criterion and each PC.

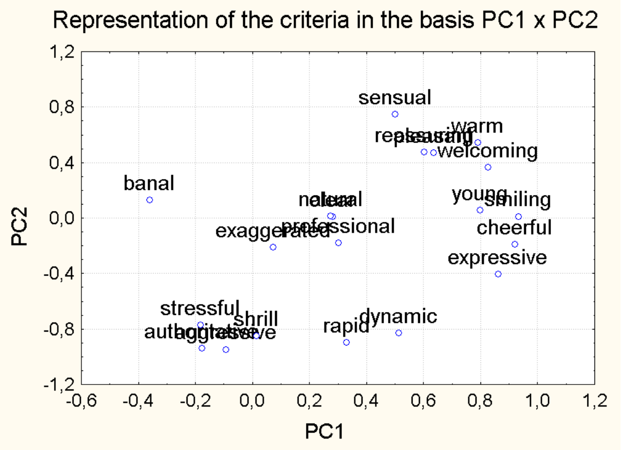

Figure 5. *Representation of the twenty subjective criteria in the PC1 x PC2 plan.*

From the study of Figure 5, PC1 might be interpreted as the young, fresh and pleasant ("young", "smiling", "cheerful", "expressive") aspects of the speech samples as opposed to a rather "banal" aspect. PC2 might be interpreted as the calm, warm and sensual ("sensual", "warm", "pleasant", "welcoming") aspects of the speech samples as opposed to the rapid and aggressive ("rapid", "dynamic", "aggressive", "authoritative") aspects. From Figure 6, PC3 might be interpreted as the classic aspect of the speech samples ("banal", "natural", "professional") as opposed to the un-classic aspect ("exaggerated"). Finally, PC4 seems difficult to interpret and might be based on a more acoustic description of the speech samples, as it is mainly correlated to the "clear" criterion.

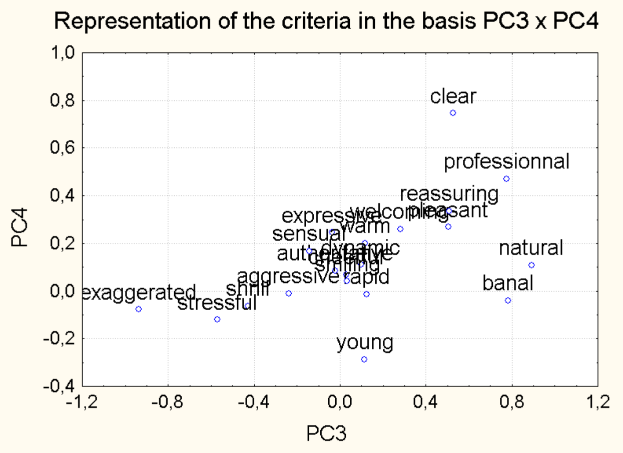

Figure 6. *Representation of the twenty subjective criteria in the PC3 x PC4 plan.*

These results tell us that listeners mainly judged speech samples on the basis of their young, smiling, cheerfulness, and expressive aspects (PC1). They then took into account the balance between the warm, calm and sensual aspects of the speech samples *versus* their rapid, stressful and aggressive aspects (PC2). Finally, to a lesser degree, they considered the classic and professional aspects of the speech samples (PC3).

At this point, it is of interest to investigate how it is possible to extract information from the speech signals that could explain the perception of listeners. If a model that connects subjective data to physical parameters could be created, it could be applied to new speech samples (new speakers or the same speakers with new styles) to give information on their perceptive portrait without running new psychoacoustic tests.

As they account for 91% of the variance of the data set, the four PCs can be used, at a first level of description, to successfully characterize the main perceived qualities of the speech samples. Nevertheless, it might be difficult to create a model since they combine several criteria that might be explained by different physical parameters or by some common physical parameters having different weights in the model of each criterion. Consequently, it was decided to elaborate twenty different models, one for each criterion.

Upon examination of the twenty criteria, it appears that some might be rather easy to model, whereas others might be very difficult. For example, a criterion such as "dynamic" could intuitively be related to the rhythm and the energy found in the signal. On the contrary, a criterion like "warm" might be much more difficult to connect to physical parameters. The following section describes the parameters that were extracted from signal analysis and subsequently correlated to the subjective data in order to create the twenty models.

4. MODEL

4.1. Extraction of Physical Parameters

Several studies have already addressed the issue of correlating physical parameters extracted from speech signals to the emotional content produced by the speaker. Scherer [8], in his comprehensive review, proposes twelve basic emotional expressions, and for each of them, gives the main related physical parameters. These parameters generally belong to three main categories: time and rhythm parameters, pitch and melodic parameters, and energy parameters. In all, two hundred and nineteen parameters were computed for each speech sample. They mostly belong to the three categories time/pitch/energy, but other parameters related to timbre were also computed. Some parameters were computed over the whole signal (*e.g.* sharpness which is the barycentre of the long-term spectrum), whereas others were computed on short time windows (pitch extracted on 30-ms windows, and energy computed on 10-ms windows). In this case, vectors are obtained for each speech sample (*e.g.*, vectors for pitch, for energy in third-octave bands). In a first step, statistical moments of the first (mean), second (variance), third (skewness), and fourth (kurtosis) order were derived from each vector in order to obtain scalar values that could be correlated to subjective data (which are defined as scalars).

4.2. First Models

After the psychoacoustic tests and the signal analysis, we have two matrixes describing the one hundred speech samples by either twenty subjective criteria, or two hundred and nineteen physical parameters. After standardization, a correlation matrix was computed from these two matrixes. For each subjective criterion, approximately twenty physical parameters show a significant absolute *r* Pearson's correlation coefficient (equal to or greater than 0.2 for the 100 observations used for the computation).

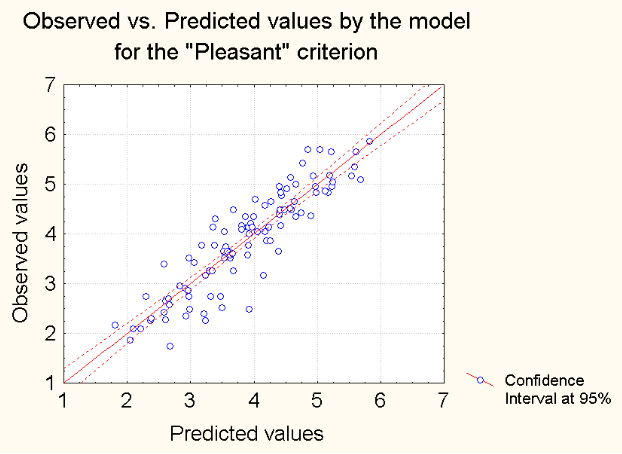

Figure 7. *Observed vs. predicted values for the "pleasant" criterion.*

For each subjective criterion, a first simple model was constructed using a linear regression algorithm: those physical parameters which were significantly correlated to it were integrated in the model as independent variables, the dependent variable needing to be explained being the subjective criterion. Twenty models were obtained, with R2 ranging from 0.6 to 0.88. Figure 7 gives an example of the observed *versus* the predicted values by the model for the criterion "pleasant". In this model, twenty-four physical parameters were included and combined linearly to predict a value for the "pleasant" criterion ranging from 1 to 7 (initial scale for the subjective tests). The R2 is equal to 0.74.

4.3. Limitations of the Model and Further Steps

The first limitation of the model is obvious: in this "death approach", nothing is known about the semantic content of the speech signal. It is known that speakers express their emotions by controlling specific acoustic parameters on specific words of the sentences they utter [9]. As a consequence, the model could be highly improved by getting information on *what* is said. Our present work focuses on this aspect, by trying to identify in several sentences the different ways of expressing the same emotion by different speakers, in order to establish a classification of the expressions of emotions for prototypical sentences.

The second limitation comes from the mathematical form of our model. In our approach, we favored a simple and explicit form that could help us to understand where to find in the signal the information that can explain the perception of listeners. It is clear that perceived emotions in speech cannot be a mere addition of several information extracted in the signal, but depend on complex non-linear processes that cannot be accounted for by a linear-regression model. Consequently, the twenty models should be highly improved by integrating non-linear approaches, including neural-networks modelling.

Finally, the time-information was lost for the most part since the time vectors derived from signal analysis were only represented by four scalars (the four first statistical moments of the vectors). Other approaches that take more fully into account this time information should also greatly improve the models.

Interestingly, speech samples that cannot be correctly modelled carry useful information that helps to identify new physical parameters to be extracted in the signal and that could better explain the listeners' judgments. For example, two speech samples were judged very differently for the "aggressive" criterion, although they had roughly the same values for the relevant physical parameters included in the "aggressive" model. This resulted in a well-predicted sample and a badly-predicted one. When listening carefully to the badly-predicted sample, it could be noticed that the articulation of the speaker was very specific, with a lot of energy present in the first syllable of most of the words of the message, leading to an unpleasant aggressive impression. As a consequence, after the analysis of this particular speech sample, a new algorithm characterizing the articulation of the speakers is now under development and it is hoped that it will significantly contribute to the explanation of the "aggressive" (and other) criterion.

5. CONCLUSIONS

This paper has proposed a methodology to model the emotional quality of speech in a telecommunication context. One hundred speech samples recorded with twenty professional speakers using five elocution styles were characterized through psychoacoustic tests by perceptive portraits on twenty subjective criteria coming from the subjects' own wordings.

Two hundred and nineteen physical parameters were extracted from each sample by signal analysis. For each criterion, the most correlated physical parameters were integrated into a linear regression model that predicts subjective data obtained in the psychoacoustic tests.

The results of this study are strongly related to the sound corpus that was recorded in a telecommunication context: the prompt and the styles used were chosen as representative of the ones that could be found in vocal servers. In a modelling perspective, this restriction to a specific context is necessary since the emotional content that can be found in speech in general is virtually infinite. Improvements of the first models proposed here are expected by taking into account more fully the semantic and time-varying information, and by developing non-linear approaches.

6. REFERENCES

[1] Yankelovich N. Levow, G.A. and Marx, M. “Designing Speech Acts: Issues in Speech User Interfaces”, in *Human Factors Computing Systems*, Proceedings of CHI95, Denver, USA, 1995, pp. 369-376.

[2] Barrass, S. *Auditory Information Design*. Ph.D. Thesis, Australian National University, 1997.

[3] Maffiolo, V. and Chateau, N. “Speech's Emotional Quality in Vocal Services”, in *Proc. Int. Conference on Affective Human Factors Design*, Singapore, June 2001, pp. 342-348.

[4] Dubois, D. *Sémantique et Cognition - Catégories, prototypes, typicalité,* 1st edition (Paris: CNRS Editions), 1991.

[5] Sattath, S. and Tversky, A. “Additive Similarity Trees”, *Psychometrika*, vol. 42, pp. 319-345, 1977.

[6] Guski, R. “Psychological Methods for Evaluating Sound Quality and Assessing Acoustic Information”, *Acustica – Acta Acustica*, vol. 83, pp. 765-774, 1997.

[7] Dilon, W.R. and Goldstein, M. *Multivariate Analysis. Methods and Applications.* John Wiley & Sons, Inc, New York, 1984.

[8] Scherer, K.R. “Vocal Affect Expression: A review and a Model for Future Research”, *Psychological Bulletin*, vol. 99(2), pp. 143-165, 1986.

[9] Rossi, M., Di Cristo, A., Hirst, D., Martin, P. and Nishinuma, Y. *L'intonation. De l'acoustique à la sémantique.* Klincksieck et Cie, Paris, France, 1981.