Prosodic Correlates of Acted *vs.* Spontaneous Discrimination of Expressive Speech: A Pilot Study

Nicolas Audibert[1,2], *Véronique Aubergé*[2,3], *Albert Rilliard*[4]

[1] Laboratoire d'Informatique d'Avignon/CERI, University of Avignon, France
[2] GIPSA Lab, UMR 5216 CNRS/Univ. Grenoble [3] LIG UMR 5217 CNRS/Univ. Grenoble, France
[4] LIMSI-CNRS, Orsay, France
nicolas.audibert@univ-avignon.fr, veronique.auberge@imag.fr, albert.rilliard@limsi.fr

Abstract

This paper presents the first results of the acoustic analysis of 12 pairs of monosyllabic acted *vs.* spontaneous expressions of satisfaction, irritation and anxiety produced by 4 subjects, discriminated and rated for emotional intensity differences in previous perceptual experiments. Acoustic features in each pair were extracted from the utterances, compared and correlated with perceptual ratings, mainly showing significant correlations between general F0 level difference in the pair and perceived emotional intensity difference, but failing to explain all the observed variability of discrimination scores. The influence of F0 contours shape of selected stimuli on perceptual discrimination scores and perceived emotional intensity is discussed.

Index Terms: expressive speech, acted emotion, spontaneous emotion, acoustic features, F0 contours

1. Introduction

Most studies on expressive speech to date have relied on utterances simulated by amateur or professional actors (see [1] for a discussion on research paradigms), since this technique is a convenient way to control the lexical contents of utterances and the recording conditions. However, the representativeness of acted expressive speech corpora for the modeling of realistic expressions of affects in speech have been questioned [2], leading to an increased effort towards collecting and analyzing spontaneous expressive corpora.

In spite of that, few studies have directly compared spontaneous vs. acted expressive speech. In a pioneering experiment, live recordings were compared to the simulation of the same recordings by professional actors [3]. From the visual inspection of resulting spectrograms, the authors concluded that acted recordings were similar enough to spontaneous ones to be considered as reliable emotional data. However, it was later shown [4] that audiovisual amusement induced by unexpected visual jokes could be perceptually discriminated from acted amusement produced by the same speakers, with a large variability in judges' discrimination performances. As the starting point of a series of cross-cultural studies, positive and negative moods were induced by inexperienced Dutch-speaking actors using Velten's technique of emotionally loaded sentences repetition, before having them simulate the same expressions [5]. Perceptual ratings of collected expressions obtained in visual-only condition indicated that acted expressions were perceived as more intense than spontaneous ones.

For the means of stimuli selection prior to a study on the typicality of vocal expressions of emotion [6], 47 acted and 146 spontaneous utterances produced by 6 actors were extracted from the E-Wiz/Sound Teacher expressive speech corpus [7], validated and rated for emotional intensity. This evaluation showed a higher perceived emotional intensity for acted utterances vs. spontaneous ones.

In a previous study [8], we extracted from the EWiz/Sound Teacher corpus 48 acted *vs.* spontaneous audiovisual expressions of satisfaction, irritation and anxiety produced by 4 semi-professional actors, and paired them by speaker and emotion. A perceptual discrimination task in audio-only, visual-only and audiovisual conditions showed that naive subjects were globally able to discriminate acted from spontaneous utterances, with a large inter-judge effect that confirms previous results on amusement [4]. Ratings of perceived emotional intensity difference in each pair were obtained using a similar protocol, revealing higher intensity ratings for acted stimuli, in line with results obtained in visual-only condition [5], and in audio-only condition for stimuli extracted from the same corpus [6]. Comparison of discrimination and intensity differences ratings showed that, although perceived intensity differences are significantly correlated to discrimination scores, they cannot account for all observed performances.

In this study, we perform an acoustic analysis of the audio stimuli that compose the 12 monosyllabic pairs, and examine the intra-pair differences in features values in light of the perceptual results. Fundamental frequency contours shapes and anchoring values are compared and discussed.

2. Acted *vs.* spontaneous expressive speech

The French audiovisual expressive corpus E-Wiz/Sound Teacher [7] was recorded using the Wizard of Oz technique, in which the subject is convinced to be interacting with a complex person-machine interface while the apparent behavior of the application is remote-controlled by the experimenter. Subjects were recruited with the pretext of participating in the last pre-commercialization tests of a novel voice-recognition-based language-learning application, presented as acting directly on subjects' brain plasticity to enable a fast and easy learning of foreign vowels pronunciation. Most of the tasks consisted in perceptual discrimination of pairs of synthetic vowels, visually presented in the acoustic triangle. The interactions of the subjects with the system were restrained to a command language composed of the French monosyllabic color names [bʁik], [ʒon], [ʁuʒ], [sabl] and [vɛʁ] and the command [paʒsʏivãt] (*next page*), enabling the collection of at least 20 utterances of each stimulus per subject, balanced across the successive phases of the scenario.

The performances attributed to the 17 subjects participating in the experiment were manipulated according to a predefined scenario. Subjects' perception skills were first

presented as among the better observed so far, prior to getting worse and worse. In the last step of the scenario, modified audio stimuli were presented to subjects to induce random choice of answers, while pretending that the learning software might have damaged their perceptual abilities. This scenario enabled the induction of both positive and negative emotions. The affects expressed were annotated by the subjects themselves from the video recording, as a first step before perceptual validation. An adapted protocol was set up for the 7 subjects who were also actors: those subjects were requested immediately after the Wizard of Oz task to express on the same utterances the affects they reported to have felt during the experiment, as well as the most frequently studied emotions (sadness, anger, fear, disgust, surprise and joy), using their acting methods. The experimenters insisted that the actors should express the affects felt in the experiment the same way they had been expressing them before. The actors recruited were practicing improvisation theater and/or street acting, and used past felt emotions as a basis for expressing emotions, as described in [9]. All of them reported the experimental set-up as optimal for applying the acting methods.

3. Background: perceptual discrimination and intensity difference ratings

Table 1. *Mean perceptual scores for the 12 monosyllabic pairs in audio-only condition. Scores range from 0 to 100. 1st value: discrimination. Scores over 50 indicate a correct discrimination. 2nd value (in parentheses): intensity difference. Scores over 50 indicate an acted stimulus rated as more intense than the spontaneous stimulus. Scores significantly over or below 50 (one-sample t-tests) are flagged with * (p<.05) or ** (p<.01).*

Speaker	Anxiety	Irritation	Satisf.	All
F1	63** (73**)	63** (64**)	43 (51)	56* (63**)
F2	58* (55*)	67** (60**)	57 (52)	61** (55**)
M1	49 (57*)	44 (46)	60* (63**)	51 (55**)
M2	56 (52)	72** (85**)	71** (72**)	66** (70**)
All	56** (59**)	62** (63**)	58** (59**)	59** (61**)

On a study focused on perceptual discrimination of acted vs. spontaneous expressive speech [8], 48 stimuli produced by two male actors (hereafter M1 and M2) and two female actors (hereafter F1 and F2) were extracted from the E-Wiz/Sound Teacher corpus [7] and paired by speaker and emotion to match every selected spontaneous stimulus with an acted stimulus. Stimuli in a pair were as much as possible selected as carrying a similar level of intensity in audio-only condition, according to the results obtained in [6]. In a first experiment, the 24 pairs were presented in a random order to 33 naïve French-speaking subjects, in audio-only, visual-only and audiovisual conditions. Subjects were requested after each pair presentation to indicate by moving a slider which one of the two presented stimuli had been uttered spontaneously by the speaker, together with their degree of confidence in this discrimination. A second experiment was set up with a similar protocol and 32 non-overlapping subjects, who were requested to indicate which stimulus expressed the most intense emotion and the extent of this difference. In each condition, each pair was presented twice, with the spontaneous stimulus presented first and with the acted stimulus presented first. In the present study, the analysis is focused on monosyllabic utterances, which enable prosodic comparison without the influence of linguistic prosody. Table 1 summarizes the scores obtained for the 12 monosyllabic pairs in audio-only condition in both perceptual evaluations.

The correlation between discrimination and intensity difference scores for those 12 pairs is significant (r=.815; p<.001), indicating that pairs with higher intensity ratings for the acted stimulus are globally better discriminated.

4. Acoustic features analysis

4.1. Methods

Acoustic features were extracted with the Praat software [10], using hand-labeled phoneme boundaries. Features used for the characterization of emotional expressions in [11] were retained, some measurements being adapted to specificities of our data. This set of features was completed with additional spectral measurements, and with voice-quality related features. Since speakers had the possibility of moving head during the recording, the mouth-to-microphone distance could not be kept constant in the whole corpus. As a result, global acoustic intensity measurements were considered as unreliable and discarded from analysis.

4.1.1. Fundamental frequency

Automatic pitch pulse detection performed by Praat autocorrelation method was hand-corrected to ensure accurate fundamental frequency measurement, following a method adapted from [12]. Mean and standard deviation of F0 were extracted together with 25th and 75th percentiles, for whole utterances and vocalic segments. F0 onset, range (defined as the difference between max and min values), and declination (defined as the difference between the end value and the start value) were also extracted as additional descriptors of F0 variations. F0 values in Hertz were converted to semitones. Cycle-to-cycle F0 perturbation measurements (jitter and shimmer) were also extracted from the vocalic segments.

4.1.2. Duration

Since this study is focused on monosyllabic utterances, speech rates measurements were limited to the duration of the utterance and to the proportion of the vocalic segment in the whole utterance duration.

4.1.3. Spectral measurements

Spectral measurements used in [11] were extracted from the long-term average spectra (LTAS) of voiced and unvoiced segments. The following parameters were extracted for the LTAS of voiced segments: the index HammI, defined as the difference between the maximum energy in the 0-2kHz band and in the 2-5kHz band; the drop-off DO1000 of spectral energy above 1kHz; the relative amount of energy in high vs. low-frequency ranges, with a cut-off frequency of 500Hz (PE500) and 1kHz (PE1000); the proportion of energy in nine non-overlapping frequency ranges from 125Hz to 8kHz. The proportion of energy in nine non-overlapping frequency

ranges from 125Hz to 8kHz was also extracted from the LTAS of unvoiced segments.
In addition of that, general descriptors of the spectral distribution were extracted on vocalic segments: spectral centre of gravity, skewness and kurtosis.
Locations and bandwidths of the first three formants were semi-automatically extracted from vowels. Due to the difficulty of performing accurate measurements, formants bandwidths were not retained in the final set of features.
Spectral measurements of voice quality parameters H1-H2 and H1-A3, respectively linked to open quotient and spectral tilt [13], were extracted from vowels and corrected for formant values location [14]. Measurement of harmonics peaks was monitored to ensure appropriate localization.

4.1.4. Intra-pair comparison

The intra-pair difference between the acted stimulus and the spontaneous one was computed for every pair and every extracted parameter. In the case of parameters calculated with a linear scale, the difference was computed as a relative proportion of acted compared to spontaneous. As for log-scale parameters, the value for the acted stimulus was subtracted from value for the spontaneous one.

Since the matching of stimuli into pairs could not be performed according to the criteria presented in section 3 while always keeping the same monosyllabic utterance, 7 pairs out of 12 had to be built using different utterances. As a result, most of the acoustic parameters extracted can be directly compared only on the 5 remaining pairs. However, since microprosodic intrinsic variations of F0 in French were shown to be very small when compared to linguistic and expressive prosodic variations [15], F0 measurements on vocalic segments can be compared across different vowels.

4.2. Results

F0-based measurements were compared across all 12 pairs, while other parameters were compared only for the 5 pairs with matching phonetic content.

Paired t-tests show a significant intra-pair difference for several F0 measurements: onset and mean on vowel ($p<.01$), standard deviation in whole utterance ($p<.01$), range in utterance ($p<.01$), 25th and 75th percentiles in utterance ($p<.05$) and in vowel ($p<.01$), while declination comparisons were non-significant, indicating a consistently higher F0 for acted vs. spontaneous stimuli. Those tests also show a significant intra-pair difference for the localization of F2 ($p<.05$), with lower values of F2 on acted stimuli, while the difference of spectral tilt is just above significance level ($p=0.05$), suggesting a more abrupt closure of vocal folds in spontaneous stimuli.

Correlations between each intra-pair difference and the ratings of emotional intensity difference were computed, showing significant correlations with F0 mean ($r=.710$; $p=.01$) and onset ($r=.617$, $p<.05$), 25th percentile ($r=.723$; $p<.01$) and 75th percentile ($r=.661$, $p<.05$) in vowel, shimmer ($r=-.927$; $p<.05$), HNR ($r=-.924$; $p<.05$), H1*-A3 ($r=-.896$; $p<.05$), proportion of spectral energy between 600 and 800 Hz in voiced segments ($r=-.887$; $p<.05$), and spectral center of gravity ($r=-.963$; $p<.01$).

No significant correlations were found between discrimination scores and intra-pair acoustic differences, except for spectral skewness ($r=.984$; $p<.01$) and kurtosis ($r=.925$; $p<.05$).

5. F0 contours comparison

Figures 1 to 4 present the fundamental frequency contours of four selected pairs, which correspond to expressions of the speakers F1 and M1 that performed the best for simulating expressions in audio-only condition according to the perceptual discrimination results presented in section 3. For each speaker, a pair discriminated over chance level (Figure 1 and Figure 3) and a pair discriminated at chance level (Figure 2 and Figure 4) were retained. F0 contours displayed on those figures are normalized to enable comparison independently of duration variations.

A first observation is that those contours confirm the tendency of higher difference in F0 general levels for acted vs. spontaneous expressions. Such tendency is in line with [16], in which activation was linked with F0 range and level while discarding the hypothesis of differentiated F0 contours shapes associated with specific emotional expressions in the data analyzed. Moreover, the comparison of contours suggests that F0 level remains higher by about 1 semitone when pairs are not discriminated over chance level. One can also note that the contour shapes of expressions of satisfaction are very similar to those previously observed on acted and spontaneous expressions of joy and satisfaction produced by speaker M2 [17].

The declination line of acted utterances, defined as the difference between attack and final F0, appears as corresponding to the declination of a declarative utterance. Indeed, the declination of French declarative utterances up to 8 syllables-long [18] corresponds to the progressive decrease of F0 related to physiological air ejection from the lungs without any laryngeal control [19].

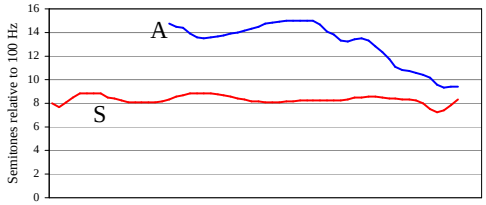

Figure 1: *F0 contours of expressions of anxiety by speaker F1. A: acted; S: spontaneous.*

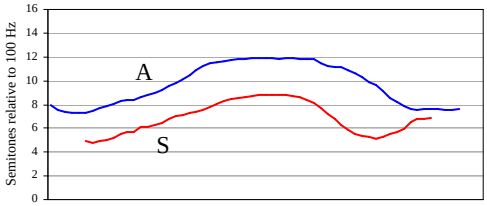

Figure 2: *F0 contours of expressions of satisfaction by speaker F1. A: acted; S: spontaneous.*

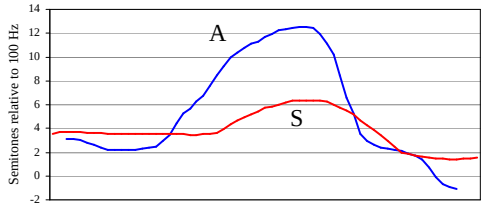

Figure 3: *F0 contours of expressions of satisfaction by speaker M1. A: acted; S: spontaneous.*

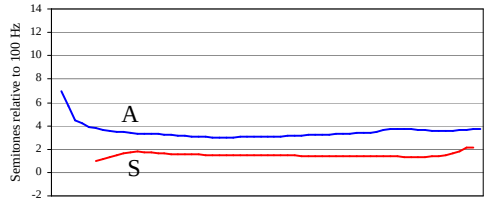

Figure 4: *F0 contours of expressions of irritation by speaker M1. A: acted; S: spontaneous.*

6. Conclusion

Intra-pair comparisons indicate that acted vs. spontaneous expressions of the same emotion by the same speaker in a very similar setting mainly differ according to their general pitch level. They also indicate that F2 is lower for acted stimuli vs. acted ones, which is an unexpected result, especially for expressions of satisfaction. One could rather expect that actors emphasize those expressions and exaggerate lips stretching, which should result in a higher F2, as noted for mechanical smiles in [20].

Correlations results suggest that pairs are better discriminated when the spectral distribution is less skewed and less sharp, according to skewness and kurtosis values. They also indicate that stimuli rated as more intense are realized with an overall higher pitch, with less perturbations and aperiodicities, and a more abrupt vocal folds closure according to shimmer, HNR and H1*-A3 values.

The comparison of F0 contours confirms the tendency of higher emotion intensity ratings associated with a larger difference in F0 general level, in line with the results obtained in [16] on the links between activation, that could be linked to perceived emotional intensity, and the level and range of F0. This comparison of F0 contours also suggests a consistently higher F0 level in acted vs. spontaneous expressions. Moreover, declination lines typical of declarative utterances observed on F0 contours of acted utterances suggests that speakers could imitate the shape of an emotional contour by means of a voluntary control, similarly to the control implied by linguistic prosody, rather than an involuntary control linked to muscular tension changes and possibly to emotion-related somatic features [21]. This imitated contour would be carried by the declination corresponding to a modal control of air pressure.

In order to further examine the role of F0 contours on perceived differences of emotional intensity, the F0 general level could be manipulated in synthesis without changing the contour shape. Such a resynthesis study will have to take into account the effects of F0 manipulation on perceived acoustic intensity, and the possible bias induced on emotional intensity ratings. Given the difficulty of collecting acted vs. spontaneous emotional expressions produced by the same speaker with a similar emotional intensity, such a study could make possible the evaluation of discrimination performances independently of emotional intensity. If the hypothesis of the link between F0 level and perceived emotional intensity is confirmed, differences in perceived intensity could indeed be neutralized using such a method. Since F0 contours of expressions of joy and satisfaction were consistently found as 'bell-shaped' in our corpus [17], those expressions being outlined as mainly carried by F0 variations in a previous resynthesis study [21], this experiment could focus on those expressions.

7. References

[1] Scherer, K. R. "Vocal communication of emotion: A review of research paradigms." Speech Comm., 40(1-2):227-256, 2003.

[2] Campbell, N. "Databases of Emotional Speech"., 1st ISCA Workshop on Speech and Emotions, Newcastle, North Ireland, 34-38, 2000.

[3] Williams, C.E., Stevens, K.N. "Emotions and speech: some acoustical correlates", Journal of the Acoustic Society of America, 52(4/2):1238-1250, 1972.

[4] Aubergé, V., and Cathiard, M. "Can we hear the prosody of smile?" Speech Communication, 40(1-2):87-97, 2003.

[5] Wilting, J., Krahmer, E., and Swerts, M. "Real vs. acted emotional speech", 10th INTERSPEECH, Pittsburgh, USA, 2006.

[6] Laukka, P., Audibert, N., and Aubergé, V. "Exploring the graded structure of vocal emotion expressions". In Hancil, S. (ed.), The Role of Prosody in Affective Speech, Linguistics insights vol. 97, Peter Lang, 241-258, 2009.

[7] Aubergé, V., Audibert, N., and Rilliard, A. "E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab". 4th LREC, Lisbon, Portugal, 179-182, 2004.

[8] Audibert, N., Aubergé, V., and Rilliard, A. "How we are not all equally competent for discriminating acted from spontaneous expressive speech" 4th Speech Prosody, Campinas, Brazil, 693-696, 2008.

[9] Enos, F., and Hirschberg, J. "A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process". 1st Int. Workshop on Corpora for Research on Emotion and Affect, Genova, Italy, 6-10, 2006.

[10] Boersma, P., and Weenink, D. Praat: doing phonetics by computer (Version 5.1.20). Retrieved from http://www.praat.org/.

[11] Banse, R., and Scherer K. R. "Acoustic Profiles in Vocal Emotion Expression". Journal of Personality and Social Psychology, 70(3):614-636, 1996.

[12] Xu, Y. "Effects of tone and focus on the formation and alignment of F0 contours". Journal of Phonetics, 27:55-105, 1999.

[13] Holmberg, E., Hillman, R., Perkell, J., Guiod, P., and Goldman, S. "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice". Journal of Speech and Hearing Research, 38:1212-1223, 1995.

[14] Hanson, H. M. Glottal characteristics of female speakers. PhD thesis, Harvard University, Division of Applied Sciences, 1995.

[15] Di Cristo, A. De la microprosodie à l'intonosyntaxe. Aix-Marseille:Université de Provence, 1985.

[16] Bänziger, T., and Scherer, K. R. "The role of intonation in emotional expressions." Speech Comm., 46:252-267, 2005.

[17] Aubergé, V., Audibert, N., and Rilliard, A. "Acoustic morphology of expressive speech: What about contours?" 2nd Speech Prosody, Nara, Japan, 91-95, 2004.

[18] Aubergé, V. *La synthèse de la parole : des règles au lexique*. PhD thesis, Université Pierre Mendès-France, Grenoble, 1991.

[19] Maeda S. A Characterization of American English Intonation. PhD thesis, Cambridge (Ma), M.I.T., 1976.

[20] Tartter, V. C., and Braun, D. "Hearing smiles and frowns in normal and whisper registers". Journal of the Acoustical Society of America, 96 (4):2101-2107, 1994.

[21] Aubergé V. "A gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP". 1st Speech Prosody. Aix, France, 151-155, 2002.

[22] Audibert, N., Vincent, D., Aubergé, V., and Rosec, O. "Expressive Speech Synthesis: Evaluation of a Voice Quality Centered Coder on the Different Acoustic Dimensions." 3rd Speech Prosody, Dresden, Germany, CD proceedings, 2006.