Acoustic Morphology of Expressive Speech: What about Contours?

Véronique Aubergé, Nicolas Audibert & Albert Rilliard

Institut de la Communication Parlée
Université Stendhal/INPG/CNRS, Grenoble, France
{auberge;audibert;rilliard}@icp.inpg.fr – http://www.icp.inpg.fr/EMOTION

Abstract

The modeling of emotional prosody in terms of contours vs. gradual cues is a recurrent question [6]. This work aims at showing that the validity of contours for characterizing emotions expressions could be revisited (1) by integrating gradual tuning on contours, in a superpositional Gestalt approach [1, 2] (2) by analyzing one parameter after the other the multiparametric contours into fine-grained contour details after having ensured that the observed stimuli express “pure” emotional variations. Acted and authentic stimuli were therefore induced following a wizard of Oz method [3]. The corpus was labeled by the selected speaker himself. The F0 contours were not a priori stylized. Analysis of “neutral” acted and authentic stimuli confirms the validity of the method and the freezing of linguistic variations in the corpus. Different kinds of pattern behaviors appeared from the analysis, with different characteristics for acted vs. authentic stimuli.

1. Introduction

Different kinds of “affects” are expressed in speech, related both on voluntary communicative controls of the speaker (i.e. linguistic strategies/expressiveness, intentional values/ attitudes), and on involuntary controls (“direct” emotions). Both in linguistics [10, 12] and in psychology [16], a discussion has been running for years about the communication levels and/or the cognitive processing involved in emotions, moods, attitudes, feeling processing… This problem is linked to the question discussed in this paper, that is the morphological structure of the expressions which carries these different affects, since these “cognitive” information are expressed together in the acoustic stream. Quite as soon as prosody became an attractive field of research, the question of the prosody of affects has been addressed. A central question is the link between the prosodic morphology of linguistic functions vs. emotional function [18, 9, 12, 13, 14]: do they share the same processing with different instantiations/parameters, or are they separate mechanisms? This implies first to adopt clear hypotheses on the morphology of linguistic prosody (i.e. to enter the complex discussion about prosodic structure, e.g. tonal vs. contour approach, contour concatenation vs. contour superposition etc.) in order to contrast with or on the contrary to follow these hypotheses for the morphology of emotional prosody; and second to describe the prosodic integration of the linguistically driven affects and the directly driven emotions.

This paper proposes to apply to the emotional function a modeling approach of prosodic morphology that we previously developed for the linguistic functions of prosody [1, 2]. It is based on the superposition of multi-parametric Gestalt contours. We introduced the notion of “gradual contour” [2] (Principle 3) either as a tuning of the superposition “weight” between carried and carrying contour or as a tuning on the multiparametric contour following the function level – now applied on the focalization function, on the F0 parameter only. The experiment presented here is the analysis of a corpus built on minimal linguistic units (monosyllabic words) aimed at freezing the realization of all linguistic functions (in particular the segmentation/hierarchization) and to represent emotional expressions only, for authentic and acted emotion variations, in the same speech acts context. The stimuli are analyzed in terms of contours frames, modulated by gradient cues. In this paper, the analysis is quite reduced to the F0 parameter, even if we claim the fundamentally multi-parametric nature of contours, in order to focalize on the discussion about F0 patterns. We try to show in these data, that some F0 contours features contribute to partially “mould” direct emotion expressions, or at least that some F0 contours features are actually used to express some emotions.

2. The nature of “expressems”

2.1. Gradual cues vs. contours characterization

Bänziger et al. [6] come back on the problem of “emotion signature in intonational patterns”. They recall that this idea, proposed earlier (for instance well established by Fonagy [10]), has been discussed and tested in parallel [18] with covariation models, implying gradual parameter variations, independently for F0 values and voice quality values. A very complete description of the gradient behavior of a large set of parameters has been given for a very large panel of emotional variations by Scherer et al. [17]. The question of the contour vs. gradual nature of the cognitive representations of expressions cannot be disconnected from the general discussion about prosodic representations. The main point to be decided is whether the processing of affective vs. other cognitive information carried by the prosodic signal are extracted/implemented following different morphological mechanisms. Rossi [15] summarizes the different theoretical approaches of prosody modeling into three categories: the superposition approach, the phonological (tonal) approach and the morphological approach (with the meaning of prosodic morphemes structures). The notion of pattern, particularly intonational pattern, in which the emotion signature can possibly be implemented, depends on the adopted theoretical approach. Only the morphological approach was clearly used for the emotion function: very early for French, Fonagy [9, 10] proposed some “melodic clichés” (global F0 contours) as vocal symbolisms (which could be related to ethological cues). Delattre [8] described some basic intonation morphemes mainly for the pragmatic level of affect. More recently, Mozziconacci [13] largely experimented the IPO approach (concatenation of minimal F0 contours or basic configurations) for varying emotional values. Ní Chasaide and Gobl [14] got some interesting results which suggest that to independently study the F0 and voice quality parameters might be a badly-formed problem for such multi-parametrical forms which reveal by inversion the vocal tract behavior.

2.2. Integration proposals

Our central hypothesis is that the perceptive separation between affective vs. linguistic treatments comes at the end of the prosodic treatment, and not just after the "parameter extraction", that is before the morphological (phonological) treatment. In this idea, the identification of the affective vs. linguistic information is precisely derived of the prosodic morphology; the prosodic analysis can decide about the nature of the encoded function. Following this hypothesis implies that (1) a cognitively relevant model of prosody is a key to identify the kind of processing (emotion vs. high level cognition) through which the information is treated after the prosodic extraction, (2) this model must be built following some morphological laws basically the same for all linguistic and non linguistic functions encoded in the prosodic signal. Like Ní Chasaide and Gobl [14], we think that the prosodic modeling must integrate the gradient and contour characterizations.

The model we have developed behind Aubergé [1] was initially based on Fonagy [10] and Delattre's [8] proposals. All the communicative functions are instantiated by Gestalt contours which do not result from concatenation (it does not belong to the morphological approach as the IPO model does). applied to the domain of the given function. The integration of the different functions is realized by a superposition of these Gestalts. Both the superposition weight and the dynamics contours themselves are tuned by gradient values, defined by the communication modeling (as for instance the focalization function [2]). The values of a given function define both the domain of the function and the domain of the prosodic encoding. The notion of domain is fundamental for affect expressions, since some of them are voluntarily controlled inside the communication exchange currency (the "pull effect" in Scherer's model, the paralinguistic affects in Léon [12], the meta and para-linguistic strategies in Aubergé [2] – attitude expressiveness) and are implemented in the domains governed by the linguistic/speech act timing. Concerning the "directly" expressed emotions (push effect in Scherer's model), the involuntary control of prosodic forms cannot be expected to be driven by the same timing laws, that is not to be "superimposed" in the linguistic domain, but coherent in time with the events causing the emotion processing.

From this point of view, to experiment the possible pattern mechanisms for direct emotion expression (the push effect of Scherer) consists first in considering the forms as multiparametrical (that means that even is we test the parameters one by one, the goal is to show a global conservation of Gestalts) and second to freeze the variations of others functions. In Bänziger and al.'s [6] perception experiment, the only variation that is explicitly asked to the professional speaker is precisely direct emotions. But a bias could be introduced by the fact that the logatome utterances are surely segmented and hierarchized by the speakers. Even in meaningless utterances, it is difficult to avoid focalization or other expressiveness cues on such long pseudo-utterances, and it is difficult on such a strong "carrying" contour to subtract the emotional expressions for which the "domain" cannot be defined, if we retain the hypothesis on a non linguistic timing for this level.

This is the reason of the method we used to build the stimuli: they must be some parts of speech as short as possible to avoid the domain problem, and they must have if possible the same speech acts values. Thus, the stimuli are mono-syllabic words, they are "answers" to implicit questions given by the task, the speakers know that they cannot influence any interlocutor turn taking, they are just in front of a software accumulating there answers.

3. Stimuli design

3.1. The induction scenario

Emotional states were induced by the perturbation of a wizard of Oz protocol, by using the Sound Teacher scenario, implemented thanks to the E-Wiz platform, specially developed for affect expression collection [3]. The Sound Teacher scenario consists in an imitation of a speech recognition-based software for the learning of foreign languages, presented to lie on the perception-action theory. The proposed learning is based on the presentation of audio stimuli (synthesized vowels) coupled with the visual presentation of associated articulator settings. The scenario is subdivided into four phases, aiming at the induction of positive, then negative, emotional states by manipulating the subjects' performances. It includes numerous perception tasks, in which the subject has to choose one vowel out of two. The answers are color names (one color per recognized sound).

Aubergé et Cathiard [4] showed that acted vs. authentic stimuli can be perceptively discriminated. In order to represent this possible opposition, some of the subjects are professional actors, first tricked with Sound Teacher in authentic performances. First of all, one of the experimenters interviewed the subjects during the debriefing occurring immediately after the recording, to investigate about the way their emotional states had been evolving along the experiment.

When the subject was an actor, immediately after the debriefing, he was asked to reproduce the emotional states he had named as experienced during Sound Teacher. He performed these "personal" labels, additionally with the classical "big six" emotions on all utterances pronounced in Sound Teacher and on semantically neutral sentences.

Afterwards, a VHS video tape of the speaker's reactions during the recording process, together with a pre-filled grid following the experiment scenario, were given to all subjects with the instruction to label the emotional states they had been feeling. One speaker, an actor, was selected for the acoustic analysis presented here. He was selected on an "expert" listening of "clear" emotional productions (before the results of the large panel perception tests in process) and because he labeled the corpus in a few simple terms, recurrent along the corpus, easy to compare with the labels he had given just after Sound Teacher which were very few. In particular he labeled several stimuli as "nothing", which could become an interesting authentic reference, that we were tempted to compare with "neutral": if the "non emotional" state is not supposed to really exist, for this subject it could result, because of the task, in the absence of vocal signs of emotion.

Thus, acted stimuli were restrained to utterances comparable to those of the spontaneous part, excluding more complex sentences also produced. 86 acted stimuli have been selected, distributed into 13 emotional labels (satisfaction, positive surprise, positive concentration, worried, anxiety, deception + neutral, joy, sadness, hot anger, disgust, fear), as well as 50 authentic stimuli recorded with Sound Teacher and distributed into 10 labels (confidence, positive concentration, joy/surprise, joy, negative concentration, deception/surprise, anxiety, anxiety/fear, weariness, nothing).

3.2. Acoustic measurements

Word and phoneme labeling of the spontaneous and acted corpus was performed thanks to the Praat software [7] by a single expert. Additionally, Praat scripts were developed to extract stimuli together with corresponding labels.

F0 contours were calculated on vowels only, located from an expert phonetic labeling. Values were extracted by means of a prosodic editor EdiProso developed at ICP and running in a Matlab environment. The F0 extraction algorithm counts, after a signal filtering, the times the signal goes down to a predefined threshold, set to 10% of amplitude for that study. Smoothed F0 contours, averaged on 32 ms frames shifted by 10 ms, were calculated from the algorithm output. Flattened contours, plotted on ten points to enable comparisons of vowels independently of duration, were also extracted.

Vowel duration values were calculated from the phonetic labeling. Those values were converted from time units to a percentage of variation around the mean (intrinsic) duration of the same vowel in the corpus, thus enabling cross-vowel comparisons. Attack and final frequency values were also extracted and used to calculate the declination line. In order to avoid calculation errors frequently occurring on signal limits, attack and final locations were shifted from 10% prior to the extraction of values. Mean and standard deviation of attack, final and duration were calculated for every emotional label.

4. Results and discussion

Table 1: Characteristic values of contours, F0 level is the difference in semitones between the attack and to the mean speaker F0, norm duration is the difference in % to 0. A emotion means acted emotion. N is negative, P positive valence, B is big, S is small arousal, as evaluated by the speaker himself

	Valence	Arousal	F0 level	F0 decl	F0dyn	norm dur
			semitones	semitones	semitones	%
A anxiety	N	B	10	-1	1	-15,9
A deception	N	S	1	1	1,5	85,6
A disgust	N	B	3	0	1	142,0
A fear	N	B	-4	6	6	14,5
A hot anger	P	B	15	3	3	29,2
A joy	P	B	11	0	1,5	16,2
A pos conc	P	S	10	-2	3	18,6
A pos surp	N	B	-2	8	8	30,2
A weariness	N	S	8	1	1	-2,9
A sadness	N	B	10	-3	3	0,4
A satisfaction	P	S	21	-3	7	77,7
A worried	N	S	0	11	11	17,9
A neutral	–	–	0	0	0,5	1,2
anxiety/fear	N	B	2	7	7	-6,6
confidence	P	S	3	-5	6	23,4
joy/surprise	P	B	-1	5	5	-12,6
weariness	N	S	-3	2	2	-14,3
neg conc	N	S	2	3	3	-20,6
nothing	–	–	0	-2	2	-14,1
pos conc	P	S	1	-4	6	-1,3
joy	P	B	1	5,5	5,5	-5,5
dec/surp	P	B	-1,5	7,5	7,5	-26,6
anxiety	N	S	1	7,5	7,5	-7,5

Table 1 presents the general characteristics of the contours. It is to be noted that the neutral contour for the acted emotions and the “nothing” contour for the authentic emotions confirms the hypothesis of the minimal intonation (reduced segmentation/hierarchisation, focalization) since the attacks of both are at the same level as the speaker’s basic vocalic F0 (which is the intonation reference in our intonation model [1, 2]; i.e. we define here, as an anchor point of contours, the F0 level which is the difference between the attack and F0 mean), the shape of the contour is flat and the declination line corresponds to the “normal” articulatory effort on such monosyllables.

The general dynamics of acted contours is lower (3,7 semi-tones) than the general dynamics of authentic contours (5,2 semi-tones). The general F0 level of acted contours is higher (6,4 semi-tones) than the authentic ones which are in average near 0 (but with significant variation). The duration of vowels (minimal rhythm) is strongly higher for acted speech (32% vs. –8,6%).

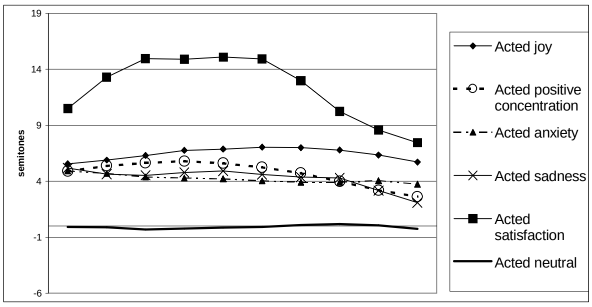

Figure 1: Acted satisfaction apart, the contours of acted joy, anxiety, sadness, pos concentration are close in form to neutral.

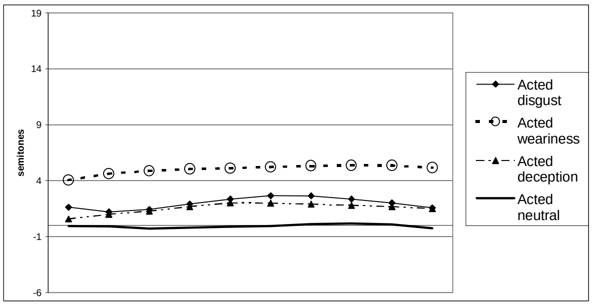

Figure 2: Acted disgust, weariness and deception have no specific prominence, but do not follow the neutral basic declination line.

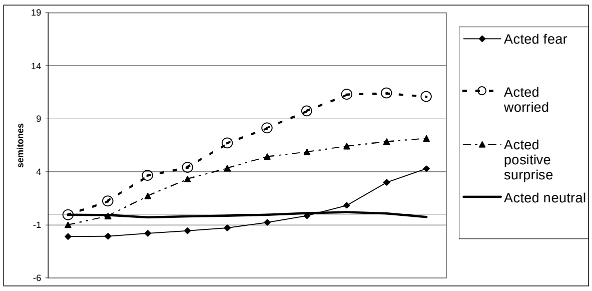

Figure 3: Acted fear, surprise and worried have similar increasing with a final prominence.

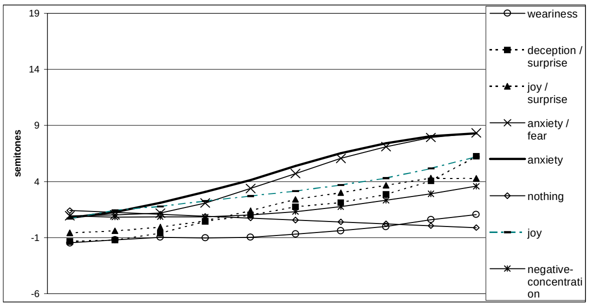

Figure 4: Authentic deception/surprise, joy/surprise, anxiety/fear, anxiety, joy on one hand, negative concentration and weariness on the other hand share similar shape cues.

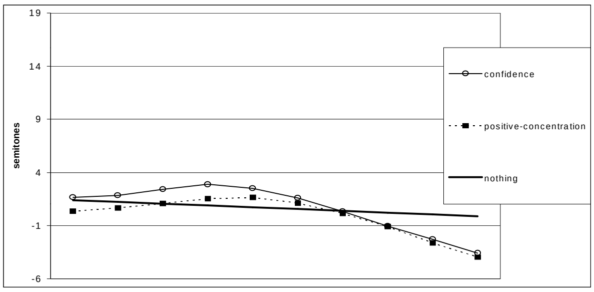

Figure 5: Authentic confidence and positive concentration have similar shape with a prominence in the first third of vowel.

The differences in shape and gradient cues (mentioned in table 1) should be interpreted as significant for expressing some cues of emotional/mental states as labeled by the speaker himself.

In parallel, we symbolized the kind of contours very roughly in 9 classical classes of contours (/ ; /¯¯; /¯ ; /\ ; _/ ; ¯¯; ¯\ ; \/ ; _; \). In parallel we measured quantitatively other parameters, as described by Scherer et al [17]. We calculated for F0 the mean, standard deviation, range, percentiles, min/max and jitter, for other source parameters the NAQ as well as 11 spectral parameters [17]. The only clear effect emerging from Anova calculation is the effect of the contour /\ on NAQ, jitter and spectral slope. But the choice of symbolic classes of contours is first not univocal to define from the dynamics of the shape and second may surely not be these classical symbols : the symbolism must include some cues which can be observed on the preceding figures. In particular the relevance of the place and threshold of glissando, psycho-acoustically validated [15] but irrelevant for linguistic prosody, could be evaluated for emotional prosody values, in particular when the timing is not linked to linguistic units.

5. Conclusion

Bänziger et al.[6] got perceptive results showing that only the overall level of F0 is relevant and confirm that general height of F0 is linkable to differential activation or arousal. These two characteristics are surely the more relevant, but we suggest that finer extra-linguistic shape cues are perhaps hidden in this experiment by "crushing" linguistic carrying contour, even when supposed to be the same for the chosen long utterances. Our aim is to replicate a similar perceptive experiment on minimal and pragmatically frozen linguistic units. In this paper we present some pre-stylizations of F0 contours, parameterized by anchor values. When compared with the references (for acted and authentic speech) different contour behaviors can be observed. We did not explain how the different labels are shared, but the contours are simple enough to test different kinds of stylizations in further perception evaluation. Some perceptive pre-tests must be held to define which kind of contour cues (as well as a possible glissando location) are sensitive to labels. Without perceptive validation, this work cannot show that contours are relevant to describe emotional speech, but it shows however that a "without emotion value" contour, for acted or authentic speech, is exactly as predicted on minimal frozen linguistic units, and that some contour variations clearly appear for emotion variations.

6. Acknowledgments

This work was held within the Japan Science and Technology Agency as part of CREST/Expressive Speech Project, directed by N. Campbell.

7. References

[1] Aubergé V., 1992. Developing a structured lexicon for synthesis of prosody, in *Talking Machine*, Bailly & Benoit Eds, Elsevier

[2] Aubergé V., 2002. A Gestalt morphology of prosody directed by functions : the example of a step by step model developed at ICP, *Proc of 1st Int Conf on Speech Prosody*, Aix-en-Provence, 151-155

[3] Aubergé, V; Audibert, N; Rilliard, A, 2003. Why and how to control emotional speech corpora. *8th European Conference on Speech Communication and Technology*, 185-188.

[4] Aubergé, V.; Cathiard, M., 2003. Can we hear the prosody of smile ? Special issue *Emotional Speech*, *Speech Communication Review* 40.

[5] Banse, R.; Scherer, K. R., 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3), 614-636.

[6] Bänziger T.; Morel M.; Scherer K. R., 2003. Is there an emotion signature in intonational patterns? And can it be used in synthesis? *Proc of Eurospeech 2003*, Geneva.

[7] Boersma P.; Weenick D.J.M, 1996. Praat, a system for doing phonetics by computer, Institute of Phonetics Sciences of Amsterdam, *Report 132*.

[8] Delattre P., 1969. L'intonation par les oppositions, *Le Français dans le Monde*, 64, Paris Hachette-Larousse, 6-13.

[9] Fonagy I., 1970. Les bases pulsionnelles de la phonation, *Revue Française de Psychanalyse*, 34, 101-136.

[10] Fonagy I., 1986..Les langages de l'émotion, *Quaterni di semantica*, 7/2, M. Alinei ed., Bologne, 305-318.

[11] Iida A.; Mokhtari P.; Campbell N., 2003. Acoustic correlates of monosyllabic utterances of Japanese in different speaking styles, *Proc. of ICPhS*, Barcelona.

[12] Léon P., 1993. *Précis de phonostylistique – parole et expressivité, Paris*, Nathan Pub.

[13] Mozziconacci S., 1995. Speeech variability and emotion: production and perception. PhD. Thesis, Technische Univeriteit Eindhoven.

[14] Ní Chasaide, A; Gobl, C.., 2003. Voice quality and expressive speech, *1st JST/CREST Int Workshop on Expressive Processing*, Kobe, 19-27.

[15] Rossi, M., 1999. Intonation : Past, Present, Future, in A.Botinis (ed.), Cambridge University Press.

[16] Scherer, K. R.; 2003. Vocal communication of emotion: A review of research paradigms, *Speech Communication Review*, (40), 227-256

[17] Scherer, K. R.; Johnston, T.; Klasmeyer, G., 2003. Vocal Expression of Emotion. In R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Eds). *Handbook of Affective Sciences*, 433-456.

[18] Scherer, K. R.; Ladd D.R.; Silverman K.E.A., 1984. Vocal cues to speaker affect: testing two models, *Journal of the Acoustic Society of America*, vol 76, 1346-1356.