Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 414–429

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

The role of voice quality and prosodic contour in affective speech perception

Ioulia Grichkovtsova [a,*], Michel Morel [a], Anne Lacheret [b,c]

[a] Laboratoire CRISCO, EA 4255, Université de Caen Basse-Normandie, Esplanade de la Paix, 14032 Caen, France
[b] UFR LLPHI, Département de Sciences du Langage, Laboratoire MODYCO, UMR CNRS 7114, Université Paris Ouest Nanterre la Défense, 200, avenue de la République, 92001 Nanterre, France
[c] Institut Universitaire de France, 103, bd Saint-Michel, 75005 Paris, France

Received 5 December 2010; received in revised form 20 September 2011; accepted 14 October 2011
Available online 25 October 2011

Abstract

We explore the usage of voice quality and prosodic contour in the identification of emotions and attitudes in French. For this purpose, we develop a corpus of affective speech based on one lexically neutral utterance and apply prosody transplantation method in our perception experiment. We apply logistic regression to analyze our categorical data and we observe differences in the identification of these two affective categories. Listeners primarily use prosodic contour in the identification of studied attitudes. Emotions are identified on the basis of voice quality and prosodic contour. However, their usage is not homogeneous within individual emotions. Depending on the stimuli, listeners may use both voice quality and prosodic contour, or privilege just one of them for the successful identification of emotions. The results of our study are discussed in view of their importance for speech synthesis.

Keywords: Speech perception; Affective prosody; Voice quality; Prosodic contour; Speech synthesis; Prosody transplantation paradigm; Attitudes; Emotions; French

1. Introduction

Affective speech attracts considerable attention of researchers both on the production and perception levels (Izdebski, 2007; Hancil, 2009). In their search for acoustic correlates of affective speech, production studies have identified the principal role of speech prosody; more precisely, such prosodic features as fundamental frequency, speech rate, intensity and voice quality are strongly correlated with affective state (Williams and Stevens, 1972; Banse and Scherer, 1996; Bachorowski, 1999; Hammerschmidt and Jürgens, 2007). Different production studies propose vocal profiles for affective states, but these profiles reveal considerable variability (Murray and Arnott, 1993; Johnstone and Scherer, 2000). No consensus exists for this phenomenon, even though some suggestions are put forward. For example, Scherer (2003) has proposed that this variation may be due to different types of arousal and intensity variation for studied affective states.

In spite of this discovered variability on the production level, perception studies show that humans can identify affective states in speech with high accuracy (Scherer et al., 2001; Pell et al., 2009b). Moreover, these studies show that affective states can be successfully identified by listeners not only in the speech expressed by speakers of the same language, but also by speakers of other languages, thus showing universality in the recognition of vocal affects similar to facial affects. Nevertheless, cross-linguistic differences are also found in affective speech perception (Elfenbein and Ambady, 2003; Pell et al., 2009a; Erickson, 2010). An important factor in cross-linguistic affective speech research may be the distinction between different types of affective states, such as emotions and attitudes.

* Corresponding author.
E-mail address: grichkovtsova@hotmail.com (I. Grichkovtsova).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.10.005

According to Scherer, emotions are usually expressed in an intense way in response to a highly significant event; in addition, the identification of emotions is largely universal (Scherer et al., 2001; Scherer, 2003). Attitudes belong to a different affective category. They may be described as relatively enduring and affectively charged beliefs and predispositions; they are less intense and more socially and culturally controlled than emotions (Scherer, 2003). More cultural differences have been observed in the identification of attitudes (Mejvaldova and Horak, 2002; Shochi et al., 2006).

In the present work, we classify perception studies of affective speech into two categories: one is focused on the level of accurate recognition of affective states in speech (Scherer et al., 2001; Dromey et al., 2005; Thompson et al., 2004), the other undertakes the perceptual evaluation of prosodic features, such as voice quality and prosodic contour, in their ability to convey affective states (Gobl and Ní Chasaide, 2003; Yanushevskaya et al., 2006; Bänziger and Scherer, 2005; Morel and Bänziger, 2004; Chen, 2005; Rodero, 2011). Here, we are principally interested in the latter one. The works of Gobl and Ní Chasaide (2003) and Yanushevskaya et al. (2006) show that voice quality alone may evoke affective associations, even though these do not exist on a one-to-one basis; a particular voice quality, for example creakiness, may be associated with several affective states. Research on the prosodic contour also shows associations between various prosodic features and affective states, even if these results vary depending on the number of prosodic features and the selection of affective states, and on the strength and nature of their associations (Yanushevskaya et al., 2006; Bänziger and Scherer, 2005; Morel and Bänziger, 2004; Chen, 2005; Rodero, 2011).

In order to avoid the influence of affective lexical meaning, perception research proposes several methodological approaches to study prosody in affective speech: speech filtering, usage of tonal sequences, usage of non-words, usage of words from foreign languages, usage of lexically neutral utterances (Lakshminarayanan et al., 2003; Thompson et al., 2004; Scherer, 2003; Bänziger and Scherer, 2005; Gobl and Ní Chasaide, 2003; Chen, 2005; Rodero, 2011). We classify speech corpora used in these studies into two main categories: utterances with or without lexical content. Perception experiments on the basis of utterances with lexical meaning establish more stable associations between affective meaning and pitch (Chen, 2005; Rodero, 2011). Perception studies on the basis of utterances without lexical meaning (Bänziger and Scherer, 2005; Yanushevskaya et al., 2006; Gobl and Ní Chasaide, 2003) show associations between prosodic contours and affective states to a certain extent, but these observations are less conclusive than those for the voice quality. We suggest that the usage of meaningless utterances may trigger a bias in the perception of the prosodic contour and minimize its importance, as it is not possible to chunk these utterances into syntactic units. Indeed, the absence of linguistic gating points for the prosodic contours may disturb the perception of affective meaning.

On the other hand, if the corpus of affective speech contains affective lexical meaning, it is difficult to fully counterbalance its influence in the perception as shown in our previous study (Grichkovtsova et al., 2007). Accentual prominence is generally realized on the informationally important words of the utterance, precisely those that are involved in the affective identification. By putting emphasis on these words, an affectively neutral prosodic contour acts as a 'lexical amplifier' and facilitates the identification of affects. These considerations motivated us to run a perception experiment on the basis of a corpus with neutral lexical meaning in order to evaluate the specific role of prosodic contour in the identification of affective states without interference with prosodic amplification for affectively charged words.

Our study is designed to investigate the perceptive value of voice quality and prosodic contour, in the identification of affective states in French. Voice quality is considered to be part of prosody (d'Alessandro, 2006). In the present work, we use the definition of the voice quality proposed by Laver (1994). It describes the voice quality through phonatory, articulatory and overall muscular tension components. We define the *prosodic contour* as the set of prosodic features other than voice quality, i.e. intensity, fundamental frequency, speech rate and rhythm. Thus, we explore the perceptual value of prosodic features not individually, but grouped under two categories, i.e. voice quality and prosodic contour.

Our main objective is to determine whether the prosodic contour and voice quality are equally important for the perception of affective states or, conversely, if one of them may be privileged. Our research hypothesis is that voice quality may be privileged for marking emotions, while variation of prosodic contour is more present in the expression of attitudes. To address our research questions, we have established a corpus of affective speech and developed an experimental methodology to evaluate the role of prosody in the identification of the affective states.

The article is organized as follows. In Section 2, we present our methodological considerations for the corpus design, our recording procedure and perception test methodology. Then, we report our data analysis based on logistic regression in Section 3. Finally, we discuss the contribution of our results to the theoretical understanding of affective speech and their implication for speech synthesis.

2. Methodology

2.1. Corpus recording

In order to build a corpus of affective speech which can be effectively used in our perception study, we considered several important methodological issues mentioned in the Introduction. A detailed methodological discussion is available in (Grichkovtsova et al., 2009). To take into account the individual variability observed in the affective speech, a multi-speaker recording was designed. We chose

not to work with professional actors who may produce stereotyped and exaggerated patterns but to record lay people through an induction method. Similar recording procedures were recently used by Murray and Arnott (2008) and Barkhuysen et al. (2010).

Twenty-two lay adult persons, native speakers of French, were recruited for the corpus recording. They were students and staff from the University of Caen. The number of male and female speakers was balanced. The recording was done with a portable recorder Microtrack 24/96 and a clip-on lavalier microphone AT831R in a quiet room. An induction technique was used in the recording procedure. A specific text was designed and tested by the authors for each studied affective state; each text contained affectively charged words and expressions appropriate for the situation. The same neutral utterance was inserted in the middle of each text: "*Je vais rentrer à la maison maintenant.*"/"*I am going home now.*" The texts are given in the Appendix. The speakers were instructed to imagine the situation described in the text and to read it three times in order to feel themselves in the described context. The idea was that the lexically neutral utterance would carry the affective state which was acted through the whole text and induced by the affectively charged text. One realization of the token utterance was extracted for each affective state to constitute the corpus. For reference, a neutral statement and a neutral interrogative were recorded according to the same induction procedure. The texts created for the recording of the neutral do not carry any emotional or attitudinal meaning, their communicative intention may be regarded as purely informational.

Two categories of affective states were recorded according to this methodology. The first category includes emotions, such as anger, disgust, fear, grief, joy, happiness and sadness. These affective states are often called basic emotions (Ekman, 1999; Izard, 1971; Plutchik, 1993). The second category represents attitudes: authority, contempt, hesitation, obviousness, politeness, shame, incredulity and surprise. The last affective state was in this category for several reasons. Even if some authors classify this affective state as an emotion (Ekman, 1999; Izard, 1971; Plutchik, 1993), others claim that it does not belong to basic emotions because it can have any valence (positive, negative or neutral) and can be combined with any emotion (Oatley and Johnson-Laird, 1987; Ortony and Turner, 1990; Power and Dalgleish, 2008). In our study, we use an induction text to record surprise with a neutral valence, and we classify it among voluntarily controlled social affects, or attitudes. Surprise has already been classified and studied as an attitude in affective speech research (Rilliard et al., 2009; Shochi et al., 2009).

2.2. Evaluation test

Our recorded corpus of lexically neutral utterances was divided into three parts and evaluated separately through the identification test. The first part was classified as emotions; it consisted of seven emotions (anger, disgust, fear, grief, joy, happiness, sadness) and a neutral statement which amounts to 176 stimuli (22 speakers × 8 states). The second part was classified as statement attitudes; it consisted of six attitudes (authority, contempt, hesitation, obviousness, politeness, embarrassment) and a neutral statement, 154 stimuli (22 speakers × 7 states). The third part was classified as doubt/interrogative attitudes; it had incredulity, surprise, neutral interrogative, and neutral statement for reference, 88 stimuli (22 speakers × 4 states). The separation of attitudes into two subgroups was motivated by the necessity to control the length of the tests; moreover, previous studies have shown that interrogative, incredulity and surprise form a perceptual category and are primarily confused between themselves (Shochi et al., 2009).

We normalized the intensity level of our stimuli according to the following procedure. First, we measured the intensities for all the neutral utterances and calculated the mean value across the 22 speakers. Next, we calculated the deviations of individual speakers in percentage for their *neutral* utterances (for example, +10% or −15% from the mean value). These calculated values were used to perform the same normalization for each affective utterance of individual speakers. Thus, all the neutral utterances have the same intensity level. This approach allows to normalize the intensity levels across speakers and to preserve individual strategies in affective realizations.

The identification test was run on a computer with the Perceval software (Ghio et al., 2007). Sennheiser headphones HD500 were used. All the stimuli were presented in a random order. The running time was about 20 min for each subtest. The listeners could listen to the stimuli just once and they had to choose one of the affective states from the list shown on the computer screen. The identification test served to validate the recorded corpus and to filter out badly identified utterances.

Twelve native French listeners (average age 30 years) participated in the test. We fixed the threshold of acceptance at 50% (at least 50% of listeners could identify the expressed affective state). This level of acceptance was chosen in order to exclude those utterances which were identified above chance but which were significantly confused with other affective states. We considered that the selection of utterances with high level of identification was important for our perception experiment. The results of this evaluation test are compiled in Table 1 which shows the number of utterances selected out of the recorded corpus with the identification level above 50%. Based on the results of the evaluation test, some affective states, such as fear, disgust, contempt, politeness and embarrassment, were finally discarded from the corpus, as they did not have enough utterances at the chosen identification level.

The selection of neutral utterances for the perception test was based on the joint analysis of the results of the three subtests. Only those utterances which were identified as neutral and scored at least 50% in all the subtests were

Table 1
The number of utterances validated for the perception test from the recorded corpus.

	Emotions	Attitudes	Neutral
	Anger – 8	Authority – 8	Statement – 8
	Grief – 5	Hesitation – 4	Question – 8
	Happiness – 5	Incredulity – 2	
	Sadness – 6	Obviousness – 3	
		Surprise – 8	
Subtotal	24	25	16

first selected. The number of validated neutral utterances was higher than the maximum validated for emotions and attitudes (8 utterances for anger, authority and surprise). In order to balance the number of validated utterances for other categories, only eight neutral utterances with the best identification level were finally selected. The number of interrogative utterances was also reduced to 8 for the same reasons.

The recording procedure was designed so that the expression of emotions and attitudes induced from the beginning of the text with emotionally marked lexical meaning would propagate in the production of the neutral utterance. Nevertheless, in practice this approach did not work for all the recorded affective states. We noticed that the realization of affective states was sometimes concentrated in one part of the text. This was the case for disgust: this emotion was mainly realized in the utterances with affectively charged words, but it did not propagate through the whole text. For our perception experiment, we were not able to use these utterances, as their lexical meaning was not neutral. The difficulty to select a high number of well identified affective utterances from our recorded corpus may be also explained by the findings reported by Barkhuysen et al. (2010). They suggested that emotions recorded through an induction method are more difficult to identify than emotions acted in a stereotyped or exaggerated way. Despite the reported difficulties, we chose the induction method in order to work with natural realizations and in perspective to apply them to speech synthesis. We are interested to see if natural realization of affective speech may be successfully used and recognized in speech synthesis.

Our induction-based methodology was designed to record a corpus of affective speech devoid of the stereotypical aspects of acted speech. By doing so we saw that it proved more difficult to obtain a recording of some affective states (fear, disgust, contempt, politeness and embarrassment) realized on the token utterance, which satisfies our *strict* criteria (high percentage of identification, and no confusion with other affective states). The problem of confusion in the identification process concerned some prosodically close affective states, i.e. obviousness/authority, and incredulity/surprise. Most of our incredulity utterances were identified as surprise, and most of our obviousness utterances were identified as authority; still we obtained several utterances according to our selection criteria. Had we discarded these two attitudes (incredulity and obviousness) more difficult for identification, this would have lowered the range of our study: the possibility to distinguish these affective states by their prosodic contour was an important issue to address.

The recording of a large number of speakers was useful to compensate, to a certain extent, for the fact that some realizations were not successful or were not actually propagated to the neutral utterance. Thus, we were able to record a corpus of 65 utterances realized by *different* speakers as shown in Table 1. The identification level of these utterances is reported in Tables 2, 4 and 7 under the 'natural' category. To summarize, only incredulity has its identification level at 50%, all the other affective states are identified above 65%. The number of selected utterances varies within affective states, but the two studied categories, emotions and attitudes, are well balanced: 24 and 25 utterances accordingly.

2.3. Perception experiment

We chose a prosody transplantation paradigm (Prudon et al., 2004) to address our main research question: what are the roles played by the voice quality and the prosodic contour in the identification of affective states? This method was previously used for the evaluation of different speech synthesizers (Garcia et al., 2006) and then used in the perception analysis of affective speech (Morel and Bänziger, 2004; Grichkovtsova et al., 2007).

The transplantation paradigm allows to extract and exchange the prosodic contour (fundamental frequency, intensity and temporal parameters) between two utterances with the same segmental content. In our case, the exchange involves natural utterances from our affective speech corpus and synthesized utterances with a neutral prosodic contour. The synthesized utterances are built with KALI, a French-speaking text-to-speech diphone synthesis system

Table 2
Percentage of correct responses for attitudes and emotions.

Affective category	Natural (%)	Prosody (%)	Voice quality (%)
Emotion	77	39	35
Attitude	71	54	13

Table 3
Output of logistic regression model for the joint analysis of emotions and attitudes. The degree of significance is coded as follows: *** for 0.0001, ** for 0.001, * for 0.01, and · for 0.05.

	Estimates β	StdErr	Wald z	Pr(>|z|)	
Intercept	0.5777	0.2250	2.567	0.0103	*
Affect	0.3176	0.1467	2.166	0.0303	*
Prosody	0.1847	0.3027	0.610	0.5418	
VQ	−3.6908	0.3588	−10.286	0.0001	***
Prosody × affect	−0.9278	0.1958	−4.739	0.0001	***
VQ × affect	0.9294	0.2189	4.246	0.0001	***

Table 4
Results of the perception test for emotions, presented in a confusion matrix. The identification level is displayed in percentages. Statistically significant values are in bold.

Encoded emotions	Type of stimuli	Decoded emotions				
		Anger	Grief	Happiness	Neutral	Sadness
Anger	Natural	**89**	0	7	4	0
	Prosody	**54**	0	14	**31**	1
	VQ	**41**	7	1	**35**	16
Grief	Natural	0	**69**	0	3	**28**
	Prosody	3	**37**	7	**27**	**26**
	VQ	11	**34**	0	17	**38**
Happiness	Natural	12	1	**77**	10	0
	Prosody	19	6	**43**	**30**	2
	VQ	12	17	11	**38**	**22**
Neutral	Natural	3	1	9	**82**	5
	Prosody	1	6	4	**77**	12
	VQ	6	15	1	**46**	**32**
	Kali	8	4	1	**80**	7
Sadness	Natural	1	19	0	12	**68**
	Prosody	19	10	0	**55**	16
	VQ	5	**32**	1	14	**48**

Table 5
Output of linear logistic regression model for emotions. The significance codes are as follows: *** for 0.0001, ** for 0.001, * for 0.01, and · for 0.05.

Coefficients	Estimates β	Std. error	z-value	Pr(>|z|)	
Intercept	0.80012	0.21622	34.700	0.0002	***
Prosody	−1.33234	0.29942	−4.450	0.0001	***
Voice quality	−1.46341	0.30218	−4.843	0.0001	***
Anger	1.26534	0.33068	3.827	0.0001	***
Happiness	0.40819	0.32127	1.271	0.2039	
Sadness	−0.03099	0.29200	−0.106	0.9155	
Prosody × anger	−0.55767	0.42124	−1.324	0.1855	
Prosody × happiness	−0.15783	0.43234	−0.365	0.7151	
Prosody × sadness	−1.10748	0.43669	−2.536	0.0112	*
VQ × anger	−0.98153	0.42405	−2.315	0.0206	*
VQ × happiness	−1.83564	0.49992	−3.672	0.0002	***
VQ × sadness	0.62759	0.40397	1.554	0.1203	

(Morel and Lacheret-Dujour, 2001). This synthesis system is used for research purposes; it is also commercialized for visually impaired people interested in high speed reading with good quality intelligibility and for industrial applications requiring high miniaturization.

Though natural speech can be used for prosody transplantation, we chose to use diphone synthesis. Diphone synthesis uses a minimal speech database containing just one instance of each diphone (sound-to-sound transition) in the language. Each diphone is selected from a corpus of dozens of diphones on the basis of such important criteria as high intelligibility and stability to prosodic variation. Diphone synthesis thus produces an unnaturally hyperarticulated voice, but it is homogeneous, intelligible and stable to changes in prosodic contours. The stability of the synthesized voice allows to concentrate the listener's attention on the prosody in perception experiments. For this reason, diphone synthesis systems, such as MBROLA, are widely used in prosody research (Dutoit et al., 1996).

Two versions of the utterance were initially synthesized with Kali: one with a male voice, the other with a female voice. These two versions were based on the realization of the utterance in an unreduced manner:

/jəvɛrãtrealamɛzõmɛ̃tənã/. As some recorded speakers used reduction in their utterances, a complementary version was synthesized to prevent difficulties in transplantation: /jvɛrãtrealamɛzõmɛ̃tnã/. The synthesized utterances do not have any vocal affective meaning, as KALI does not perform this type of treatment. Nevertheless, we include the original synthesized utterances in the perception experiment to see if they are associated with any affective state.

We used our validated corpus of 65 utterances for the preparation of the audio stimuli (24 utterances for emotions, 25 utterances for attitudes and 16 utterances for neutral). In the process of prosody transplantation, the prosodic contour of Kali was mapped onto the utterances encoded by the speakers, and the prosodic contours of

the speakers were mapped onto the utterances synthesized with Kali. Three versions of each utterance were thus developed: version 1 – 'natural' (natural prosodic contour and voice quality encoded by the speaker), version 2 – 'voice quality' (natural voice quality of the speaker and prosodic contour from Kali), version 3 – 'prosody' (natural prosodic contour of the speaker and voice quality from Kali). For control reasons, we included original synthesized utterances in the audio stimuli (2 versions for the male voice and for the female voice). It was important to experimentally confirm that synthesized utterances were identified as neutral, and that the prosodic contour of Kali does not carry any vocal affective meaning. In total, we designed 199 audio stimuli: 65 'natural' stimuli, 65 'voice quality' stimuli, 65 'prosody' stimuli and 4 pure Kali stimuli.

2.3.1. Effect of prosody transplantation on voice quality

The term of voice quality has two definitions. Laver (1980, 1994) proposed this term to describe the total vocal image of a speaker. In this broad sense, voice quality refers to phonatory, articulatory and overall muscular tension of individual speech. It is also used in a narrow sense, for example by Gobl and Ní Chasaide (2003) and Campbell and Mokhtari (2006), to describe only phonation types or laryngeal qualities, such as breathiness, creakiness, harshness.

In our study, we used the prosody transplantation to create our audio stimuli. The prosodic contour was exchanged between a natural utterance and a synthesized utterance. It is supposed that the prosody transplantation acts on the prosodic contour, i.e. its fundamental frequency, intensity and temporal parameters, but that it does not modify the voice quality characteristics, its phonatory, articulatory and tension components, as defined by Laver (1994). Thus, in principal, the prosody transplantation allows to study the effect of prosodic contour without voice quality modification of the transformed signal. In reality, we could observe that this method worked quite effectively for neutral speech, while for affective speech, separation between voice quality and prosodic contour was sometimes difficult due to their interaction. These difficulties concerned especially those affective utterances for which speakers used important modifications of articulatory reduction level, overall tension or pitch excursions.

The articulation rate[1] is one of the prosodic features which interacts strongly with voice quality parameters, e.g. the articulatory characteristics of speech. The articulatory range describes the excursion size of movements for the lips, jaw and tongue. When the articulation rate increases, the speakers reduce their articulation range. As a result, vowels are less distinguished from each other, some of them are reduced or even dropped; consonants are also reduced, assimilated or dropped. Conversely, when the articulation rate decreases, the articulatory range gets larger: speakers produce larger contrasts between vowels and more precise articulation of consonants. Moreover, the articulation range is often correlated with a particular overall muscular tension: laxness participates in the narrow articulation range, while tenseness is associated with wider ranges of labial, mandibular and lingual articulation. Pitch parameters also interact with voice quality. For example, wider pitch excursions are associated with more tenseness and vowel prolongation.

During the transplantation process, a natural affective utterance receives a neutral prosodic contour. The articulation rate and pitch parameters of the new prosodic contour can sometimes considerably contradict the original articulatory reduction level, muscular tension and articulation range, resulting in an undesirable phenomenon: an impression of unnaturally high speech rate, excessive muscular tension, etc. Hence, some of the 'voice quality' stimuli may be perceived as atypical or unnatural.

On the contrary, the transfer of the natural prosodic contour to an utterance produced by a diphone synthesizer does not cause as much degradation as for the natural stimuli. In this type of synthesis, each diphone is selected for its acoustic quality and flexibility, and normalized in order to match perfectly to its neighbor diphones. As a result, the vocal material of the synthesized utterance is much more flexible than natural speech, and it is very intelligible in spite of an impression of some artificiality. This is the reason why diphonen synthesis is often used in perception experiments on prosodic contours, including the quality evaluation of these contours (Garcia et al., 2006; Prudon et al., 2004; Morel and Bänziger, 2004). Nevertheless, more degradation was observed in the 'prosody' stimuli, when the natural prosodic contour contained extremely high pitch excursions.

2.3.2. Perception test method

The perception experiment was run as two subtests. The first subtest dealt with emotions: anger, grief, sadness, happiness and neutral statement (100 stimuli[2]). The second subtest dealt with attitudes: authority, hesitation, incredulity, obviousness, surprise, neutral statement and neutral interrogative (127 stimuli[3]). Twenty native French speakers (11 females and 9 males) participated in the test. They were students and staff of the University of Caen, France. Their average age was 27 years old.

The test was run on a computer with the Perceval software (Ghio et al., 2007). Good quality Sennheiser headphones HD500 were used. All the stimuli were presented in a random order. The running time was about 25 min for each subtest; participants had a pause between the two subtests. The listeners could listen to the stimuli just

[1] The rate at which a given utterance is produced (Laver, 1994).

[2] Number of stimuli: [24 utterances (emotions) + 8 utterances (neutral statement)] × 3 (types) + 4 utterances (Kali) = 100.

[3] Number of stimuli: [25 utterances (attitudes) + 16 utterances (neutral statement and interrogative)] × 3 (types) + 4 utterances (Kali) = 127.

once and they had to choose one of the affective states from the list shown on the computer screen. Four types of stimuli were included in the test: 'natural' (natural prosodic contour and voice quality encoded by the actor), 'voice quality' (natural voice quality of the speaker and prosodic contour from Kali), 'prosody' (natural prosodic contour of the speaker and voice quality from Kali), Kali (prosodic contour and voice quality of Kali). As the perception experiment uses one single lexically neutral utterance, the successful identification relies entirely on the prosodic characteristics of the used stimuli.

3. Statistical analysis and results

The results of the identification test form a set of categorical data, which are reported in confusion matrices (Table 4 for emotions and Table 7 for attitudes). We transformed the response variable into a binary indicator of whether an affective state encoded in the vocal stimulus is successfully identified or not (0 – not identified, 1 – successfully identified). Logistic regression is an appropriate statistical tool for categorical data analysis, see Baayen (2004), Jaeger (2008) and Beck and Vasishth (2009) for detailed discussions. We performed our analysis using R statistical software (R Development Core Team, 2008) to examine the role of voice quality and prosodic contour in the identification of the studied affective states. We analyzed the two affective categories jointly and separately, and did an individual analysis of each affective state afterwards.

3.1. Joint analysis of emotions and attitudes

Our first analysis concerns the whole set of studied affective states, both emotions and attitudes: (`anger`, `grief`, `happiness`, `sadness`, `authority`, `hesitation`, `incredulity`, `obviousness`, `surprise`). The percentage of correct responses for attitudes and emotions is summarized in Table 2. Responses to `neutral` stimuli were *not* included in the analysis, as they do not belong to the affective categories. The dependent variable is an indicator of whether a stimulus was successfully identified or not. We consider a logistic regression model for the probability π_i that stimulus i is successfully identified, with type of stimulus and affective category as explanatory variables. The model has the following form:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ \text{Prosody}_i + \beta_2\ \text{VQ}_i + \beta_3\ \text{Affect}_i + \beta_4\ \text{Prosody}_i \times \text{Affect} + \beta_5\ \text{VQ}_i \times \text{Affect}_i \quad (1)$$

'Prosody' and 'VQ' (voice quality) are dummy variables for type of stimulus, taking 'Natural' as the reference. 'Affect' is a dummy variable for affective category (0 – emotion, 1 – attitude). We also include the interaction between type of stimulus and affective category (Prosody × Affect, VQ × Affect). The analyzed data set has the total of 2940 responses. The model formula for the R function *lrm* is given in Eq. (1).

As shown in Table 3, the regression reveals that the probability of a successful identification for attitudes is higher than for emotions ($\beta = 0.3176$, StdErr $= 0.1467$, Wald $z = 2.166$, $p = 0.0303$). The probability to identify 'prosody' stimuli is less than 'natural' stimuli, but the difference is not significant ($\beta = 0.1847$, StdErr $= 0.3027$, Wald $z = 0.610$, $p = 0.5418$). The probability to identify 'voice quality' stimuli is significantly smaller ($\beta = -3.6908$, StdErr $= 0.3588$, Wald $z = -10.286$, $p = 0.0001$), the negative value of β indicates that for 'voice quality' stimuli, the probability of being not identified is higher than that of being identified. The interaction between the type of stimuli and the affective categories is significant: Prosody × Affect: $\beta = -0.9278$, StdErr $= 0.1958$, Wald $z = -4.739$, $p = 0.0001$; VQ × Affect: $\beta = 0.9294$, StdErr $= 0.2189$, Wald $z = 4.246$, $p = 0.0001$. It shows that successful identification differs between attitudes and emotions for different types of stimuli.

3.2. Emotions

Our second statistical analysis deals with emotions. Identification results for emotions are given in Fig. 1 and Table 4. As defined in Section 2.3.2, four types of stimuli are represented: 'natural', 'voice quality', 'prosody' and 'Kali'. In the present work we consider 5 affective categories (`anger`, `grief`, `happiness`, `sadness` and `neutral statement`). So the chance level is $100\% \div 5 = 20\%$. The results show that 'Kali' stimuli were identified as `neutral` at

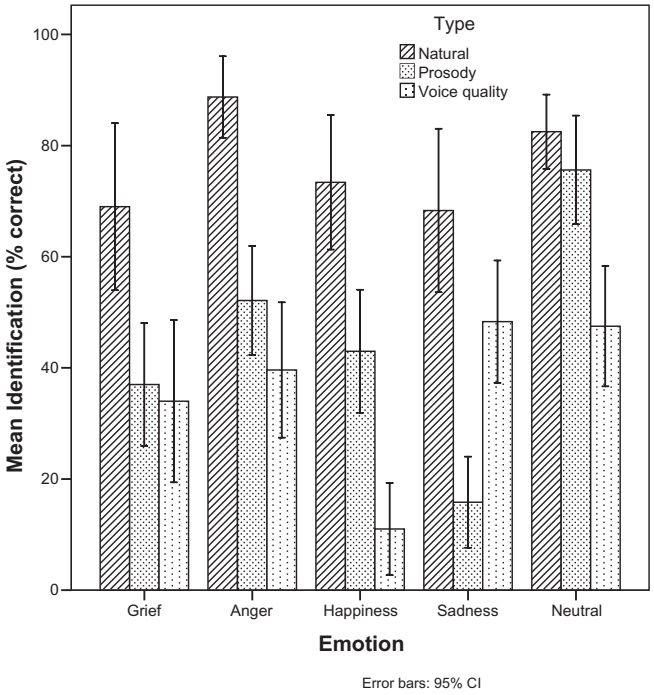

Fig. 1. Identification of emotions in percentage with confidence intervals 95%.

80% which considerably exceeds the chance level. It confirms that the prosodic contour used in the design of 'voice quality' stimuli is not affectively charged, and successful identification of 'voice quality' stimuli cannot be based on the prosodic contour. The identification results for 'Kali' stimuli are reported only in Table 4 and are not included in the subsequent statistical tests.

Perception test results are presented through a confusion matrix in Table 4. Encoded emotions are plotted against the decoded emotions. This allows to study the percentages of correct responses and the pattern of confusions. Results that surpass the chance level of 20% and whose confidence intervals are over the chance level are considered as statistically significant; they are displayed in bold.

The regression analysis for emotions includes `anger`, `grief`, `happiness` and `sadness`. Responses to `neutral` stimuli were *not* included in the analysis, as they do not belong to the affective category. The dependent variable is an indicator of whether a stimulus was successfully identified or not. We consider a logistic regression model for the probability π_i that stimulus i is successfully identified, with type of stimulus and type of emotion as explanatory variables. The model has the following form:

$$\begin{aligned}\log\left(\frac{\pi_i}{1-\pi_i}\right) = {} & \beta_0 + \beta_1\ \text{Prosody}_i + \beta_2\ \text{VQ}_i \\ & + \beta_3\ \text{Anger}_i + \beta_4\ \text{Happiness}_i \\ & + \beta_5\ \text{Sadness}_i + \beta_6\ \text{Prosody}_i \times \text{Anger}_i \\ & + \beta_7\ \text{Prosody}_i \times \text{Happiness}_i \\ & + \beta_8\ \text{Prosody}_i \times \text{Sadness}_i \\ & + \beta_9\ \text{VQ}_i \times \text{Anger}_i \\ & + \beta_{10}\ \text{VQ}_i \times \text{Happiness}_i \\ & + \beta_{11}\ \text{VQ}_i \times \text{Sadness}_i \end{aligned} \quad (2)$$

'Prosody' and 'VQ' (voice quality) are dummy variables for type of stimulus, taking 'Natural' as the reference. 'Anger', 'Happiness', 'Sadness' are dummy variables for type of emotion, taking 'Grief' as the reference. We also include the interaction between type of stimulus and type of emotion. The analyzed data set represents a total of 1440 responses. The model formula for the R function *lrm* is given in Eq. (2). The results are presented in Table 5.

The logistic regression results are significant for 'prosody' ($\beta = -1.33234$, StdErr $= 0.29942$, Wald $z = -4.450$, $p = 0.0001$) and 'voice quality' ($\beta = -1.46341$, StdErr $= 0.30218$, Wald $z = -4.843$, $p = 0.0001$). These results confirm that there are differences in the identification level between types of stimuli: both 'prosody' and 'voice quality' types are less identified than 'natural'. We also notice that the coefficients for 'prosody' and 'voice quality' variables are very close and we additionally test whether the successful identification differs for these two types of stimuli. The null hypothesis for this test is that the coefficients of 'prosody' and 'voice quality' are equal:

$$\beta_1 - \beta_2 = 0 : \text{null hypothesis} \quad (3)$$

To carry out this test we loaded the additional *car* library and used *linearHypothesis* R function. The result is not significant ($\chi^2 = 0.1964$, $p = 0.6576$). Therefore, we conclude that 'prosody' and 'voice quality' types do not differ in their chance of being successfully identified. Both voice quality ($M = 35\%$) and prosodic contour (39%) contribute to the identification of emotions, and these contributions are equally important.

The regression also reveals that `anger` has significantly better identification than other emotions ($\beta = 1.26534$, StdErr $= 0.33068$, Wald $z = 3.827$, $p = 0.00013$). It also shows the interaction between some emotions and types of stimuli. More precisely, `sadness` has the least identification of 'prosody' stimuli; `anger` and `happiness` are statistically less identified in 'voice quality' stimuli than `grief` and `sadness`.

We performed logistic regressions for individual emotions and neutral according to the formula (4) and report their results in Table 6. The dependent variable is an indicator of whether a stimulus was successfully identified or not. We consider a logistic regression model for the probability π_i that stimulus i is successfully identified, with type of stimulus as an explanatory variable. 'Prosody' and 'VQ' are dummy variables for type of stimulus, taking 'Natural' as the reference.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ \text{Prosody}_i + \beta_2\ \text{VQ}_i \quad (4)$$

`Anger` results show that 'natural' stimuli have the highest identification ($M = 89\%$), 'prosody' ($M = 54\%$) and 'voice quality' ($M = 41\%$) stimuli have statistically lower identification (their coefficients are in Table 6). We also tested the coefficients for 'prosody' and 'voice quality' variables to see whether their identification differs. The test was done assuming the null hypothesis (3) and using the *linearHypothesis* R function. We obtained $\chi^2 = 6.0261$ and $p = 0.0141$, which means that they are significantly different. 'Natural' stimuli are identified only as `anger`, 31% of 'prosody' and 35% of 'voice quality' stimuli are identified as `neutral`.

For `grief`, the regression results are significant: 'Natural' stimuli ($M = 69\%$) have the highest identification level, they differ from the other two types of stimuli, see Table 6. 'Prosody' ($M = 37\%$) and 'voice quality' ($M = 34\%$) accuracy results are not significantly different from each other ($\chi^2 = 0.1964$, $p = 0.6576$). Confusion has occured with the `sadness` category: 28% of 'natural' stimuli, 26% of 'prosody' stimuli and 38% of 'voice quality' stimuli were identified as `sadness`.

`Happiness` results are also significant. The three types of stimuli are significantly different from each other: 'natural' ($M = 77\%$) has a higher identification level than 'prosody' ($M = 43\%$) and 'voice quality' ($M = 11\%$), see Table 6. 'Prosody' and 'voice quality' accuracy results are significantly different from each other ($\chi^2 = 22.891$, $p = 0.0001$). `Happiness` is the only emotion whose 'voice quality' stimuli are not identified. We suppose that smiling, which is present in 'natural' stimuli of `happiness`, is not

Table 6
Output of linear logistic regression model for individual emotions. The degree of significance is coded as follows: *** for 0.0001, ** for 0.001, * for 0.01, and · for 0.05.

Emotion	Coefficients	Estimates β	Std. error	z-value	Pr(>|z|)	
Anger	Intercept	2.0655	0.2502	8.256	0.0001	***
	Prosody	−1.8900	0.2963	−6.379	0.0001	***
	Voice quality	−2.4449	0.2975	−8.218	0.0001	***
Grief	Intercept	0.8001	0.2162	3.700	0.000215	***
	Prosody	−1.3323	0.2994	−4.450	0.0001	***
	Voice quality	−1.4634	0.3022	−4.843	0.0001	***
Happiness	Intercept	1.2212	0.2373	5.147	0.0001	***
	Prosody	−1.5031	0.3116	−4.824	0.0001	***
	Voice quality	−3.3120	0.3980	−8.321	0.0001	***
Sadness	Intercept	0.7691	0.1962	3.919	0.0001	***
	Prosody	−2.4398	0.3179	−7.675	0.0001	***
	Voice quality	−0.8358	0.2681	−3.118	0.00182	**
Neutral	Intercept	1.5506	0.2081	7.453	0.0001	***
	Prosody	−0.4184	0.2778	−1.506	0.132	
	Voice quality	−1.6507	0.2614	−6.314	0.0001	***

preserved in 'voice quality' stimuli by the transplantation method. Previous research (Aubergé and Cathiard, 2003) has shown that listeners can reliably identify smiling in speech and use it in recognition of affective states. Thus, if smiling is possibly damaged by the prosody transplantation, 'voice quality' results for `happiness` must be treated carefully. Confusion results show that `happiness` has 30% of 'prosody' and 38% of 'voice quality' stimuli identified as `neutral`.

For `sadness`, the three types of stimuli are also significantly different from each other. 'Natural' stimuli have the highest identification ($M = 68\%$), and it is statistically significant, see Table 6. 'Voice quality' stimuli are identified with a lower accuracy ($M = 48\%$). 'Prosody' stimuli ($M = 16\%$) are not identified as `sadness`, and they differ significantly from 'voice quality' stimuli ($\chi^2 = 26.827$, $p = 0.0001$). Sadness is the only emotion whose 'prosody' stimuli were not identified. Confusion results for `sadness` show that 55% of the 'prosody' stimuli are identified as `neutral`, suggesting a similarity of the prosodic contours of `neutral` and `sadness`. The 'voice quality' stimuli are confused with `grief` at 32%. These identification results for 'voice quality' stimuli highlight the role of voice quality in the identification of `sadness`.

Results are significant for `neutral`. 'Natural' ($M = 82\%$) and 'prosody' ($M = 77\%$) stimuli are not significantly different from each other, see Table 6. 'Voice quality' stimuli are identified at a lower level ($M = 46\%$), they are significantly different from 'natural', see Table 6, and from 'prosody' stimuli ($\chi^2 = 25.753$, $p = 0.0001$). Only 'voice quality' stimuli are identified as another category; 32% of these stimuli are associated with `sadness`.

3.3. Attitudes

The data for attitudes are illustrated in Fig. 2 and Table 7. Three types of stimuli are represented: 'natural' (natural prosodic contour and voice quality encoded by the speaker), 'voice quality' (natural voice quality of the speaker and prosodic contour from Kali) and 'prosody' (natural prosodic contour of the speaker and voice quality from Kali). We consider seven affective categories (`authority`, `hesitation`, `incredulity`, `obviousness`, `surprise`, `neutral statement` and `neutral interrogative`), so the chance level is 14%. Table 7 also includes the results for 'Kali' stimuli (prosodic contour and voice quality of Kali) to examine if the Kali prosodic con-

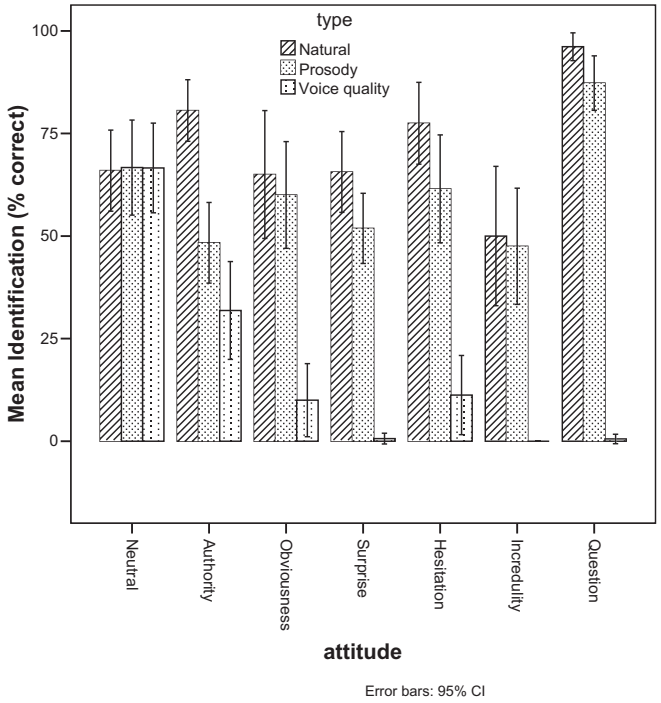

Fig. 2. Identification of attitudes in percentage with confidence intervals 95%.

Table 7
Results of the perception test for attitudes, presented in a confusion matrix. The identification level is displayed in percentages, statistically significant values are in bold. Abbreviations: Aut – authority, Hes – hesitation, Inc – incredibility, Neu – neutral, Obv – Obviousness, Int – interrogative, Sur – surprise.

Encoded attitudes	Type of stimuli	Decoded attitudes						
		Aut	Hes	Inc	Neu	Obv	Int	Sur
Authority	Natural	**81**	0	0	12	5	0	2
	Prosody	**51**	0	1	**39**	9	0	0
	VQ	**32**	4	1	**48**	14	1	0
Hesitation	Natural	0	**77**	4	9	6	4	0
	Prosody	6	**63**	5	11	10	3	2
	VQ	9	11	1	**69**	10	0	0
Incredulity	Natural	0	5	**50**	0	0	7	**38**
	Prosody	0	5	**48**	0	0	7	**40**
	VQ	**40**	0	0	**43**	12	0	5
Interrogative	Natural	0	0	1	1	0	**96**	2
	Prosody	0	4	2	2	0	**87**	5
	VQ	**22**	7	0	**53**	**17**	1	0
Neutral	Natural	9	1	1	**66**	**23**	0	0
	Prosody	9	2	1	**67**	**21**	0	0
	VQ	10	9	2	**67**	12	0	0
	Kali	15	1	1	**71**	11	1	0
Obviousness	Natural	10	3	0	13	**65**	7	2
	Prosody	12	3	4	**18**	**60**	0	3
	VQ	**27**	10	0	**52**	10	0	1
Surprise	Natural	0	1	**21**	0	0	12	**66**
	Prosody	1	3	**21**	1	0	**22**	**52**
	VQ	**37**	4	2	**43**	12	1	1

tour carries any affective information. The results show that 'Kali' stimuli were identified as `neutral` statements at 71%, which considerably exceeds the chance level (14% correct). It confirms our previous results for emotions and shows that the prosodic contour used in the design of 'voice quality' stimuli is not associated with attitudes. Successful identification of 'voice quality' stimuli cannot be based on the prosodic contour. The identification results for 'Kali' stimuli are reported only in Table 7 and are not included in the subsequent statistical tests.

We use a confusion matrix in Table 7 to present perception test results. Encoded attitudes are plotted against the decoded attitudes. This allows to study the percentages of correct responses and the pattern of confusions. Results that surpass the chance level of 14% and whose confidence intervals are over the chance level are considered as statistically significant; they are displayed in bold.

The regression analysis for attitudes includes `authority`, `hesitation`, `incredulity`, `obviousness` and `surprise`. Responses to `neutral` stimuli (statement and interrogative) were *not* included in the analysis, as they do not belong to the affective category. The dependent variable is an indicator of whether a stimulus was successfully identified or not. We consider a logistic regression model for the probability π_i that stimulus i is successfully identified, with type of stimulus and type of attitude as explanatory variables:

$$\begin{aligned}\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1\ \text{Prosody}_i + \beta_2\ \text{VQ}_i \\ &+ \beta_3\ \text{Obviousness}_i + \beta_4\ \text{Surprise}_i \\ &+ \beta_5\ \text{Hesitation}_i + \beta_6\ \text{Incredulity}_i \\ &+ \beta_7\ \text{Prosody}_i \times \text{Obviousness}_i \\ &+ \beta_8\ \text{Prosody}_i \times \text{Surprise}_i \\ &+ \beta_9\ \text{Prosody}_i \times \text{Hesitation}_i \\ &+ \beta_{10}\ \text{Prosody}_i \times \text{Incredulity}_i \\ &+ \beta_{11}\ \text{VQ}_i \times \text{Obviousness}_i \\ &+ \beta_{12}\ \text{VQ}_i \times \text{Surprise}_i \\ &+ \beta_{13}\ \text{VQ}_i \times \text{Hesitation}_i \\ &+ \beta_{14}\ \text{VQ}_i \times \text{Incredulity}_i \end{aligned} \quad (5)$$

'Prosody' and 'VQ' are dummy variables for type of stimulus, taking 'Natural' as the reference. 'Hesitation', 'Incredulity', 'Obviousness' are dummy variables for type of attitude, taking 'Authority' as the reference. We also include the interaction between type of stimulus and type of attitude. The analyzed data set has a total of 1500 responses. The model formula for the R function *lrm* is given in Eq. (5), and the results are presented in Table 8.

The logistic regression results are significant for 'prosody' ($\beta = -1.4008$, StdErr $= 0.2550$, Wald $z = -5.494$, $p = 0.0001$) and 'voice quality' ($\beta = -2.1853$, Std-

Table 8
Output of linear logistic regression model for attitudes. The degree of significance is coded as follows: *** for 0.0001, ** for 0.001, * for 0.01, and · for 0.05.

Coefficients	Estimates β	StdErr	Wald z	Pr(>|z|)	
Intercept	1.4258	0.2000	7.128	0.0001	***
Prosody	−1.4008	0.2550	−5.494	0.0001	***
Voice quality	−2.1853	0.2623	−8.332	0.0001	***
Obviousness	−0.8068	0.3366	−2.397	0.0001	*
Surprise	−0.7792	0.2602	−2.994	0.0027	**
Hesitation	−0.1891	0.3342	−0.566	0.571595	
Incredulity	−1.4258	0.3742	−3.811	0.0001	***
Prosody × obviousness	1.1872	0.4558	2.605	0.0092	**
Prosody × surprise	0.8292	0.3432	2.417	0.0157	*
Prosody × hesitation	0.6749	0.4359	1.548	0.1216	
Prosody × incredulity	1.3007	0.5150	2.526	0.0115	*
VQ × obviousness	−0.6309	0.5720	−1.103	0.2701	
VQ × surprise	−3.5302	1.0501	−3.362	0.0008	***
VQ × hesitation	−1.1169	0.5154	−2.167	0.0302	*
VQ × incredulity	−14.3807	379.4016	−0.038	0.9697	

Err = 0.2623, Wald $z = -8.332$, $p = 0.0001$). These results confirm that there are differences in the identification level between types of stimuli: both 'prosody' and 'voice quality' types are less identified than 'natural'. We also tested whether the successful identification differs for these two types of stimuli and found that the difference is significant ($\chi^2 = 11.443$, $p = 0.0007$). Thus, 'natural' stimuli have the best identification level ($M = 71\%$), 'prosody' stimuli are also well identified but at a lower identification level ($M = 54\%$). 'Voice quality' stimuli were not successfully identified ($M = 13\%$). These results reveal that the prosodic contour plays the decisive part in the identification of attitudes.

The regression also reveals that `authority` has significantly better identification than other attitudes, see Table 8. It also shows the interaction between some attitudes and types of stimuli, more precisely, `authority` and `hesitation` have higher identification of 'prosody' stimuli than other attitudes; `surprise` and `incredulity` have the lowest identification of the 'voice quality' stimuli.

We performed logistic regressions for individual attitudes, as well as for two neutral categories used in the experiment, neutral statement and interrogative. The regression was performed according to the same formula as that used for emotions, Eq. (4). These results are reported in Table 9.

For `authority`, the results are significant. The three types of stimuli are significantly identified. 'Natural' stimuli have the highest identification ($M = 81\%$); they are significantly different from the other two types of stimuli, see Table 9. 'Prosody' ($M = 51\%$) and 'voice quality' ($M = 32\%$) differ significantly from each other ($\chi^2 = 11.443$, $p = 0.0007$). `Authority` is the only attitude whose 'voice quality' category was significantly identified. The confusion matrix shows that, for `authority`, 39% of 'prosody' and 48% of 'voice quality' stimuli are identified as `neutral`.

`Hesitation` results are significant. 'Natural' ($M = 77\%$) and 'prosody' ($M = 63\%$) stimuli are identified, and they are significantly different from each other, ($\beta = -0.7259$, StdErr = 0.3536, Wald $z = -2.053$, $p = 0.0401$). 'Natural' and 'prosody' stimuli are not significantly confused with other attitudes. The 'voice quality' stimuli are significantly different from 'prosody' stimuli ($\chi^2 = 37.178$, $p = 0.0001$). 'Voice quality' stimuli are not successfully identified as `hesitation` ($M = 11\%$), but as `neutral` at 69%.

For `incredulity`, the results are significant. 'Natural' ($M = 50\%$) and 'prosody' ($M = 48\%$) stimuli are identified, they are not significantly different from each other ($\beta = -0.02500$, StdErr = 0.09239, Wald $z = -0.271$, $p = 0.787$). The confusion matrix in Table 7 shows that 38% of 'natural' and 40% of 'prosody' stimuli are identified as `surprise`. The 'voice quality' stimuli differ from 'natural' stimuli ($\beta = -0.50000$, StdErr = 0.09239, Wald $z = -5.412$, $p = 0.0001$) and from 'prosody' stimuli ($\chi^2 = 26.431$, $p = 0.0001$); they are not identified as `incredulity` ($M = 0\%$), but as `neutral` at 43% and as `authority` at 40%.

`Obviousness` results are significant. The identification accuracy for 'natural' ($M = 65\%$) and 'prosody' ($M = 60\%$) is statistically similar with ($\beta = -0.2136$, StdErr = 0.3778, Wald $z = -0.565$, $p = 0.5718$). The 'voice quality' stimuli are not successfully identified ($M = 10\%$), and they are found to be significantly different from 'natural' stimuli ($\beta = -2.8163$, StdErr = 0.5083, Wald $z = -5.540$, $p = 0.0001$) and from 'prosody' stimuli ($\chi^2 = 26.607$, $p = 0.0001$). The confusion matrix shows that 18% of 'prosody' stimuli are identified as `neutral`. 'Voice quality' stimuli are identified at 52% as `neutral` and at 27% as `authority`.

Results are also significant for `surprise`. 'Natural' ($M = 66\%$) and 'prosody' ($M = 52\%$) stimuli are successfully identified, but they are significantly different from each other ($\beta = -0.5716$, StdErr = 0.2297, Wald $z = -2.489$, $p = 0.0128$). Some confusion is observed for these two types of stimuli: 21% of 'natural' and 'prosody' stimuli are identified as `incredulity`, another 22% of 'prosody' stimuli are identified as `interrogative`. The 'voice quality' stimuli are not identified as `surprise` ($M = 1\%$), and they are

Table 9
Output of linear logistic regression models for individual attitudes. The significance codes are as follows: *** for 0.0001, ** for 0.001, * for 0.01, and · for 0.05.

Emotion	Coefficients	Estimates β	StdErr	Wald z	Pr(>|z|)	
Authority	Intercept	1.4258	0.2000	7.128	0.0001	***
	Prosody	−1.4008	0.2550	−5.494	0.0001	***
	Voice quality	−2.1853	0.2623	−8.332	0.0001	***
Hesitation	Intercept	1.2368	0.2677	4.619	0.0001	***
	Prosody	−0.7259	0.3536	−2.053	0.0401	*
	Voice quality	−3.3022	0.4437	−7.442	0.0001	***
Incredulity	Intercept	0.50000	0.06533	7.653	0.0001	***
	Prosody	−0.02500	0.09239	−0.271	0.787	
	Voice quality	−0.50000	0.09239	−5.412	0.0001	***
Obviousness	Intercept	0.6190	0.2707	2.287	0.0222	*
	Prosody	−0.2136	0.3778	−0.565	0.5718	
	Voice quality	−2.8163	0.5083	−5.540	0.0001	***
Surprise	Intercept	0.6466	0.1665	3.885	0.0001	***
	Prosody	−0.5716	0.2297	−2.489	0.0128	*
	Voice quality	−5.7155	1.0166	−5.622	0.0001	***
Neutral	Intercept	1.5506	0.2081	7.453	0.0001	***
	Prosody	0.02808	0.23699	0.118	0.906	
	Voice quality	0.02808	0.23699	0.118	0.906	
Interrogative	Intercept	3.2452	0.4161	7.799	0.0001	***
	Prosody	−1.2993	0.4799	−2.707	0.00678	**
	Voice quality	−8.4326	1.0853	−7.770	0.0001	***

significantly different from 'natural' stimuli ($\beta = -5.7155$, StdErr = 1.0166, Wald $z = -5.622$, $p = 0.0001$) and from 'prosody' stimuli ($\chi^2 = 25.668$, $p = 0.0001$). The 'voice quality' stimuli are identified as `neutral` at 43% and as `authority` at 37%.

`Interrogative` results are significant. 'Natural' ($M = 96\%$) and 'prosody' ($M = 87\%$) stimuli are successfully identified only as `interrogative`; they are significantly different from each other, see Table 9. 'Voice quality' stimuli differ from 'prosody' stimuli ($\chi^2 = 47.921$, $p = 0.0001$); they are not identified ($M = 1\%$) as `interrogative`, but as `neutral` ($M = 53\%$), `authority` ($M = 22\%$) and `obviousness` ($M = 17\%$).

For `neutral statement`, the results show that 'natural' ($M = 66\%$), 'prosody' ($M = 67\%$) and 'voice quality' ($M = 67\%$) are successfully identified, but they are not significantly different from each other, Table 9. Confusion results show that 23% of 'natural' and 21% of 'prosody' stimuli are identified as `obviousness`, and the 'voice quality' results are not significantly confused with other categories.

`Incredulity`, `interrogative`, `hesitation`, `obviousness` and `surprise` have the same identification pattern: 'natural' and 'prosody' stimuli are successfully identified with similar or slightly different identification accuracy, while 'voice quality' stimuli are not identified. `Authority` differs from the other affective states in the category of attitudes: both 'prosody' and 'voice quality' stimuli are significantly identified, even if their level of identification is lower than that for 'natural' stimuli. The matrix of confusion shows that 'voice quality' stimuli of all the studied affective states are mainly confused with `neutral`. 'Prosody' stimuli of `authority` and `obviousness` were also confused to some extent with `neutral`. `Surprise` and `incredulity` have some of their 'prosody' stimuli confused between themselves.

3.4. Individual strategies within affective states

In our statistical analysis of emotions based on logistic regression, we found that both voice quality and prosodic contour were important in the successful identification of stimuli. The number of utterances per affective state (5–8 utterances) did not allow us to test utterances as a random factor in our logistic regression models. Indeed, it is recommended to have at least 30 groups for a random factor to produce valid estimates for multi-level logistic regression models (Moineddin et al., 2007). Larger samples are particularly recommended for the analysis of binomial data (Hox, 2002). Thus, we performed only a descriptive statistical analysis to examine to what extent the value of voice quality and prosodic contour vary in the identification of individual utterances for the same emotions. This analysis also included `authority` which was the only attitude with successfully identified 'voice quality' stimuli. The identification results are presented in Tables 10 and 11. We used a multi-speaker corpus in our perception experiment, thus each utterance represents a different speaker. The results show variability in the identification of individual utterances, thus suggesting that depending on the utterance and hence on the speaker, different strategies are used in the identification of the same affective state. The prosodic

Table 10
Identification results for individual stimuli of emotions.

Stimuli	Natural	Prosody	VQ
anger12.wav	100	95	75
anger8.wav	100	85	25
anger5.wav	95	50	50
anger19.wav	95	35	55
anger17.wav	85	25	30
anger13.wav	80	60	30
anger15.wav	80	20	25
anger9.wav	75	65	35
grief3.wav	80	10	45
grief6.wav	75	15	30
grief1.wav	65	80	40
grief5.wav	65	40	10
grief7.wav	60	40	45
happiness2.wav	95	65	20
happiness13.wav	85	55	10
happiness14.wav	70	30	25
happiness22.wav	70	45	0
happiness3.wav	65	20	0
sadness10.wav	80	45	45
sadness12.wav	75	10	60
sadness14.wav	70	30	50
sadness8.wav	65	0	45
sadness3.wav	60	0	45
sadness6.wav	60	10	45

Table 11
Identification results for individual stimuli of authority.

Stimuli	Natural	Prosody	VQ
authority19.wav	100	65	40
authority9.wav	95	90	15
authority14.wav	95	25	30
authority3.wav	85	100	65
authority17.wav	75	5	30
authority12.wav	70	55	20
authority11.wav	65	15	25
authority5.wav	60	50	30

contour and the voice quality may be equally important in the identification, as for utterances recorded in the files *anger12.wav*, *anger5.wav* and *grief7.wav*. The prosodic contour may be privileged for utterances as *anger8.wav* and *grief1.wav*. The voice quality may be privileged as in utterances *anger19.wav* and *grief3.wav*.

Another important observation from the individual analysis is that high identification level is possible for utterances with different strategies. For example, we can compare the identification of two stimuli *anger12.wav* and *anger8.wav* which are both identified at 100% in the natural version. The first stimuli has the strategy based on the joint usage of voice quality and prosody. The second stimuli relies mostly on the prosodic contour. The effective expression of affective states may thus be achieved by a number of possible strategies which may either privilege voice quality or prosodic contour, or use them in combination.

In spite of some transplantation difficulties (see supra happiness in 3.2), our results show that prosodic contour and voice quality are used in the identification of the studied affective states. Moreover, the present analysis shows that the perceptive role of prosodic contour and voice quality may vary for individual utterances within the same affective state.

4. Discussion and conclusions

A series of perception tests was performed to study the role of voice quality and prosodic contour in the identification of affective states in French. We initially hypothesized that voice quality may be privileged in the perception of emotions, while the identification of attitudes may rely on the prosodic contour. This hypothesis is confirmed by our results only partly. The prosodic contour is involved not only in the perception of attitudes, but also in the perception of emotions. Voice quality is mainly associated with emotions. Authority is the only attitude in our corpus successfully identified on the 'voice quality' stimuli. Hence, the difference between attitudes and emotions examined in this experiment is the following: the perception of attitudes is based mainly on the prosodic contour, while emotions rely on both voice quality and prosodic contour. These results support previous studies (Gobl and Ní Chasaide, 2003; Yanushevskaya et al., 2006) showing that voice quality may be associated with emotions. It also shows that attitudes may be identified by the prosodic contour in French. Moreover, our experiment also suggests that emotions may be successfully distinguished on the basis of prosodic contours. We consider the last finding as especially important, because several perception studies report difficulties to show associations between prosodic features and emotions (Yanushevskaya et al., 2006; Bänziger and Scherer, 2005; Morel and Bänziger, 2004). However, we should highlight that our study differs on several methodological points. First, we did not separate the prosodic contour into its individual components, but studied it globally, as a set of prosodic features (intensity, fundamental frequency, speech rate and rhythm). Second, we used a natural rather than a stylized prosodic contour. Third, our corpus was recorded on a neutral utterance with lexical meaning. These factors could have contributed to the successful identification of affective states on the basis of the prosodic contour in our experiment. The question stands what results may be obtained with this methodology in other languages, if other languages have the same value of prosody in the identification of affective states as in French. Further research is required to extend the present study to other languages and thus explore universal and cross-linguistic features in the perception of affective states.

Our perception experiment was based on prosody transplantation paradigm. The transplantation method was initially designed to evaluate neutral synthesized speech. The application of this method to affective speech proved useful. It allowed to uncover interesting aspects of the role of prosody in the identification of affective states: both attitudes and emotions can be distinguished by their specific

prosodic contour. Nevertheless, we observed some limits in the usage of prosody transplantation for affective speech. The first difficulty concerned prosodic contours with high pitch excursions, as they could not be reproduced in the synthesized speech in a natural way without any changes in the voice quality parameters. Secondly, we noticed that transplantation paradigm did not work equally well for the voice quality of different emotions. For example, the voice quality of anger was better preserved than that of happiness. These difficulties put forward some correlations between the prosodic contour and the voice quality. Important modifications of the prosodic contour may require changes in the voice quality, and vice versa.

In spite of the imperfect reproduction of the natural voice quality in our synthesized stimuli, emotions give clearly more importance to the voice quality in their identification than attitudes do. As the transfer of a neutral prosodic contour to a natural affective utterance does not fully preserve the original voice quality, 'voice quality' results must be treated carefully. In reality, the role of voice quality may be even more important. Nevertheless, in spite of these difficulties, we obtain coherent results for 'voice quality' stimuli proving that some of the voice quality information has been transmitted. Indeed, the transplantation method does not preserve the voice quality and the prosodic contour to the full extent. However, this methodology allows to make new advances in the study of affective prosody. More specifically, it shows the importance of the prosodic contour in the identification of affective states.

The analysis of individual utterances gives an additional insight in the prosody transplantation method and its usage in affective speech research. Our interest to study individual utterances was initially motivated by the possibility to examine the level of variability between individual speakers. According to our recording methodology, individual utterances represent different speakers. The group results have shown that both voice quality and prosodic contour are used in the identification of emotions.

The analysis of individual utterances has allowed to discover that several strategies are possible in the identification of the same emotion. Listeners may either rely on both voice quality and prosody, or privilege one of them. Moreover, a high identification level may be observed independently from the chosen strategy for individual utterances. An utterance relying on the prosodic contour can have the same high identification as an utterance using a joint combination of voice quality and prosodic contour. It attests the possibility for speakers to express affective speech effectively by different strategies. This observation not only contributes to our theoretical understanding of affective speech expression, but it also presents an interest for affective speech synthesis, as some strategies may be easier for replication in the speech synthesis than others. Voice quality is more difficult to modify in speech synthesis than prosodic contours, especially in real time. The possibility to extract prosodic contours from natural speech and to apply them in affective speech synthesis is of interest. Though most diphone synthesizers use stylized prosodic contours, see Schröder (2008) for a detailed review, there is certainly an interest in the usage of natural prosodic contours to increase the naturalness in affective speech synthesis.

Our study reports new results on the usage of voice quality and prosodic contour in affective speech perception. We recorded and used a multi-speaker corpus in our experiment. This approach proved useful as we discovered various successful strategies in affective speech. The number of utterances representing studied affective states does not allow to perform an exhaustive analysis of individual strategies. Further, it would be interesting to pursue our analysis on the production level and identify the role of individual prosodic features in different strategies; work along these lines is in progress. In addition, our experimental approach may be used in cross-linguistic research with non-native speakers of French or with participants who have no knowledge of French to discover universal and language specific characteristics of affective speech.

Acknowledgments

We acknowledge the financial support from the Conseil Régional de Basse-Normandie, under the research funding program "Transfer de Technologie". I.G. is pleased to thank Dr. Henni Ouerdane for careful reading of the manuscript, useful discussions and assistance with statistics.

Appendix A. Texts used for the recording of affective states

A.1. Emotions

Anger: Vous appelez ça une chambre d'hôtel? Regardez un peu ces draps! Ils sont ignobles. Vous ne croyez quand même pas que je vais dormir ici! C'est révoltant! *Je vais rentrer à la maison maintenant!* Ce n'est pas un hôtel ici, c'est un élevage de cafards! (Do you call this a hotel room? Look at these sheets! You don't think that I am going to sleep here! It is disgusting! *I am going home now!* It is not a hotel here; it is a cockroach farm.)

Grief: Tu sais comme j'aimais mon chien? Hier, quand je suis revenu(e) de voyage, j'ai appris qu'il était mort. Je suis bouleversé(e). *Je vais rentrer à la maison maintenant.* J'vais jamais pouvoir le dire aux enfants. (You know how I loved my dog, don't you? Yesterday when I came back from my trip I found out that she had died. I am overwhelmed. *I am going home now.* How will I be able to tell the children?)

Happiness: Mon frère reviendra demain! Quelle joie! Je suis si content(e)! *Je vais rentrer à la maison maintenant!* Je vais annoncer cette super nouvelle à ma famille! (My brother is coming tomorrow! Such a joy! I am so happy! *I am going home now!* I will tell this excellent news to my family!)

Sadness: Ce que tu m'as appris m'a fichu le moral à zéro. C'est vraiment déprimant... *Je vais rentrer à la maison maintenant.* J'ai l'impression que c'est une situation sans

issue. (Your news makes me feel so low. It is so depressing. *I am going home now.* I have the impression that there is no way out.)

A.2. Attitudes

Authority: Vous devez finir ce projet absolement! N'oubliez pas de poster aussi ces lettres! *Je vais rentrer à la maison maintenant.* Je vous appellerai pour voir où vous en êtes. Dépêchez-vous! C'est vraiment urgent! (You must finish the project imperatively tonight. Don't forget to also send these letters. *I am going home now.* I will call to see how it is going. Hurry up! It is urgent!)

Incredulity: C'est vraiment le patron qui m'a demandé de partir? *Je vais rentrer à la maison maintenant?* Ça m'étonnerait. Il y a juste une heure, il m'a demandé de rester ici ce soir. (Is it really the boss who said this? *I am going home now?* It's strange. Just one hour ago he asked me to stay here tonight.)

Hesitation: Il est cinq heures de l'après-midi? Peut-être je vais rester encore un peu. Bien que... Peut-être... *Je vais rentrer à la maison maintenant.* Bon, je ne suis pas encore sûr. (Is it five p.m.? I may stay a bit more. Though... Maybe... *I am going home now.* Well, I am not sure yet.)

Obviousness: Il est plus de 20 heures. Je dois partir, tu sais bien. *Je vais rentrer à la maison maintenant.* Mes amis doivent m'attendre. (It is past 8 p.p. I must leave, you know. *I am going home now.* My friends must be waiting for me.)

Surprise: Quoi? Qu'est-ce que tu dis? *Je vais rentrer à la maison maintenant?* C'est une blague? (What? What are you saying? *I am going home now?* Is it a joke?)

A.3. Neutral

Neutral statement: J'ai fini de ranger les boites. Elles sont classées et numérotées. *Je vais rentrer à la maison maintenant.* Je reviendrai demain à dix heures. (I have finished putting the projects in order. They are filed and numbered. *I am going home now.* I will be back tomorrow at ten in the morning.)

Neutral interrogative: Tous les dossiers sont prêts. Vous n'avez plus besoin de moi? *Je vais rentrer à la maison maintenant?* Vous pouvez m'appeler, s'il y a un problème. (All the projects are ready. *I am going home now?* You can call if you need me.)

References

Aubergé, V., Cathiard, M., 2003. Can we hear the prosody of smile? Speech Comm. 40, 87–97.

Baayen, R.H., 2004. Statistics in psycholinguistics: a critique of some current gold standards. Mental Lexicon Working Papers 1, 1–45.

Bachorowski, J.A., 1999. Vocal expression and perception of emotion. Curr. Dir. Psychol. Sci. 8, 53–57.

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70, 614–636.

Bänziger, T., Scherer, K.R., 2005. The role of intonation in emotional expressions. Speech Comm. 46, 252–267.

Barkhuysen, P., Krahmer, E., Swerts, M., 2010. Crossmodal and incremental perception of audiovisual cues to emotional speech. Lang. Speech 53, 3–30.

Beck, S., Vasishth, S., 2009. Multiple focus. J. Semantics 26, 159–184.

Campbell, N., Mokhtari, P., 2006. Voice quality: the 4th prosodic dimension. In: Proc. XVth Internat. Congress of Phonetic Sciences, Barcelona, Spain, pp. 2417–2420.

Chen, A.J., 2005. Universal and language-specific perception of paralinguistic intonational meaning. Ph.D. Thesis.

d'Alessandro, C., 2006. Voice source parameters and prosodic analysis. In: Sudhoff, S., Lenertová, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N., Schließer, J. (Eds.), Methods in Empirical Prosody Research. De Gruyter, Berlin, New York, pp. 63–87.

Dromey, C., Silveira, J., Sandor, P., 2005. Recognition of affective prosody by speakers of English as a first or foreign language. Speech Comm. 47, 351–359.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., van der Vrecken, O., 1996. The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In: ICSLP, Philadelphia, pp. 1393–1396.

Ekman, P., 1999. Basic emotions. In: Dalgleish, T., Power, T. (Eds.), The Handbook of Cognition and Emotion. The Guilford Press, New York.

Elfenbein, H.A., Ambady, N., 2003. Universal and cultural differences in recognizing emotions. Curr. Dir. Psychol. Sci. 12, 159–164.

Erickson, D., 2010. Perception by Japanese, Korean and American listeners to a Korean speaker's recollection of past emotional events: some acoustic cues. In: Speech Prosody 2010, Chicago.

Garcia, M.N., d'Alessandro, C., Bailly, G., de Mareuil, P.B., Morel, M., 2006. A joint prosody evaluation of French text-to-speech systems. In: Proc. LREC, pp. 307–310.

Ghio, A., André, C., Teston, B., Cavé, C., 2007. PERCEVAL: une station automatisée de tests de PERCeption et d'EVALuation auditive et visuelle. TIPA 22, 115–133.

Gobl, C., Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Comm. 40, 189–212.

Grichkovtsova, I., Lacheret, A., Morel, M., 2007. The role of intonation and voice quality in the affective speech perception. In: Proc. Interspeech.

Grichkovtsova, I., Morel, M., Lacheret, A., 2009. Perception of affective prosody in natural and synthesized speech: which methodological approach? In: Hancil, S. (Ed.), The Role of Prosody in Affective Speech. Peter Lang, pp. 371–390.

Hammerschmidt, K., Jürgens, U., 2007. Acoustical correlates of affective prosody. J. Voice 21, 531–540.

Hancil, S. (Ed.), 2009. The Role of Prosody in Affective Speech. Peter Lang.

Hox, J., 2002. Multilevel Analysis: Techniques and Applications. Lawrence Erlbaum Associates, Mahwah, NJ.

Izard, C., 1971. The Face of Emotion. Appleton-Century-Crofts, New York.

Izdebski, K., 2007. Emotions of the Human Voice. In: Vol. 1–3. Plural Publishing.

Jaeger, T.F., 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. Mem. Lang. 59, 434–446.

Johnstone, T., Scherer, K.R., 2000. Vocal communication of emotion. In: Lewis, M., Haviland, J.M. (Eds.), Handbook of Emotions. Prentice Hall, New Jersey, pp. 220–235.

Lakshminarayanan, K., Shalom, D.B., van Wassenhove, V., Orbelo, D., Houde, J., Poeppela, D., 2003. The effect of spectral manipulations on the identification of affective and linguistic prosody. Brain Lang. 84, 250–263.

Laver, J., 1980. The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge.

Laver, J., 1994. Principles of Phonetics. Cambridge University Press, Cambridge.

Mejvaldova, J., Horak, P., 2002. Synonymie et homonymie attitudinale en tchèque et en français. In: Proc. Internat. Conf. on Speech Prosody 2002, Aix-en-Provence, France.

Moineddin, R., Matheson, F.I., Glazier, R.H., 2007. A simulation study of sample size for multilevel logistic regression models. BMC Med. Res. Methodol. 7, 1–45.

Morel, M., Bänziger, T., 2004. Le rôle de l'intonation dans la communication vocale des émotions: test par la synthèse. Cah. Inst. Ling. Louvain 30, 207–232.

Morel, M., Lacheret-Dujour, A., 2001. Kali, synthèse vocale à partir du texte: de la conception à la mise en oeuvre. Trait. Automat. Lang. 42, 1–29.

Murray, I.R., Arnott, J.L., 1993. Towards the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Amer. 93, 1097–1108.

Murray, I.R., Arnott, J.L., 2008. Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. Comput. Speech Lang. 22, 107–129.

Oatley, K., Johnson-Laird, P.N., 1987. Towards a cognitive theory of emotions. Cognition Emotion 1, 29–35.

Ortony, A., Turner, T.J., 1990. What's basic about basic emotions? Cognition Emotion 97, 315–331.

Pell, M.D., Monetta, L., Paulmann, S., Kotz, S.A., 2009a. Recognizing emotions in a foreign language. J. Nonverb. Behav. 33, 107–120.

Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., Kotz, S.A., 2009b. Factors in the recognition of vocally expressed emotions: a comparison of four languages. J. Phonetics 37, 417–435.

Plutchik, R., 1993. Emotions and their vicissitudes: emotions and psychopathology. In: Lewis, M., Haviland, J.M. (Eds.), Handbook of Emotions. John Wiley and Sons, Ltd.

Power, M., Dalgleish, T., 2008. Cognition and Emotion: From Order to Disorder. Psychology Press.

Prudon, R., d'Alessandro, C., Boula de Mareuil, P., 2004. Unit selection synthesis of prosody: evaluation using diphone transplantation. In: Narayanan, S., Alwan, A. (Eds.), Text-to-speech Synthesis: New Paradigms and Advances. Prentice Hall, New Jersey, pp. 203–217.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rilliard, A., Shochi, T., Martin, J.C., Erickson, D., Aubergé, V., 2009. Multimodal indices to Japanese and French prosodically expressed social affects. Lang. Speech 52, 223–243.

Rodero, E., 2011. Intonation and emotion: influence of pitch levels and contour type on creating emotions. J. Voice 25, 25–34.

Scherer, K.R., 2003. Vocal communication of emotions: a review of research paradigms. Speech Comm. 40, 227–256.

Scherer, K.R., Banse, R., Wallbott, H.G., 2001. Emotion inferences from vocal expression correlate across languages and cultures. J. Cross-Cult. Psychol. 32, 76–92.

Schröder, M., 2008. Approaches to emotional expressivity in synthetic speech. In: Izdebski, K. (Ed.), Emotions in the Human Voice: Culture and Perception. Plural Publishing, pp. 307–321.

Shochi, T., Aubergé, V., Rillard, A., 2006. How prosodic attitudes can be false friends: Japanese vs. French social affects. In: Proc. Internat. Conf. on Speech Prosody 2006, Dresden, Germany.

Shochi, T., Rilliard, A., Aubergé, V., Erickson, D., 2009. Intercultural perception of social affective prosody. In: Hancil, S. (Ed.), The Role of Prosody in Affective Speech. Peter Lang, pp. 31–60.

Thompson, W.F., Schellenberg, E.G., Husain, G., 2004. Decoding speech prosody: do music lessons help? Emotion 4, 46–64.

Williams, C.E., Stevens, K.N., 1972. Emotions and speech: some acoustic correlates. J. Acoust. Soc. Amer. 52, 1238–1250.

Yanushevskaya, I., Gobl, C., Ni Chasaide, A., 2006. Mapping voice to affect: Japanese listeners. In: Proc. 3rd Internat. Conf. on Speech Prosody 2006, Dresden, Germany.