AN ACOUSTIC AND PERCEPTUAL STUDY ON THE EMOTIVE SPEECH IN KOREAN AND FRENCH

CHUNG Soo-Jin
Institut de Phonétique, 19 rue de Bernardins Paris, France

ABSTRACT

This study presents the acoustic and perceptual analysis of the emotive speech loaded with anger, joy, sorrow, or tenderness, in Korean and French. Statistic analysis found the factors which affected the identification of emotions, such as emotion-type, comprehension level of a given language, modality of phrase, etc. Based on the acoustic similarities, we regrouped the studied emotions into active vs. passive emotion group and positive vs. negative emotion group. The perceptual confusion of the emotions in a same group was explained mainly by the similarity of the activity aspect. We reported also the acoustic filtering experience which had an effect on the identification of emotions in relation to the mother tongue of listeners.

1. INTRODUCTION

The emotion is a complex phenomenon ; a given emotion is considered as a result of the interaction between acoustic, physiological, and psychological features.

In the speech analysis, the personal or emotional aspect was somewhat ignored in contrast with the rich literature of the lexical or grammatical aspect. In this paper, we review briefly the general prosody and the notion of emotion, and make a acoustic and perceptual study of the vocal expression of emotions, such as anger, joy, sorrow, and tenderness, using neutral sentences as a reference.

2. GENERAL CONCEPT

Every prosodic feature seems to have a motivated origin, ethological, physiological or psychological, and this paralinguistic origin may explain the similar prosodic tendencies through non-related languages, for example, Korean and French.

In general, the biological necessity to breathe at regular interval creates the pause which can be also present for marking different degrees of frontier. The downward pitch contour is due to the depression of subglottal pressure and the gesture for a sentence or a meaning group can be described as a set of tension and relaxation of articulators.

Concerning the perception, one does not hear directly physical variations. Fraisse (1974) suggested two perceptual organizations, "Accentual rhythm" of initial segment and "Temporal rhythm" related to the final lengthening. The coexistence of two rhythms exercises contradictory forces on the interpretation of rhythm.

These general tendencies can be modified by semantic or emotional emphasis.

Many of terms used to describe emotions, not being clearly defined in the literature. Because it is impossible to quantify the emotion and there is no objective rules to define emotive terms.

According to Scherer (1981), the emotion is "the organism's interface to the world outside", having three principal functions ; "a) they reflect the evaluation of the relevance and significance of particular stimuli, b) they physiologically and psychologically prepare the organism for appropriate action, c) they communicate the organism's state and behavioral intentions to other organisms in the surroundings". He noted also that emotion is not a steady state condition, but a process of events, which arise in rapid succession following a stimulus event ; a given emotion is the result of a series of "stimulus evaluation checks" in the "component patterning model".

Many emotion theories use the concept of "basic" emotions but there is few agreement as to what constitutes a "basic emotion". In the combinational emotion theory called a "palette theory" by Scherer, new emotions are produced by mixing the primary basic emotions together. In terms of universality, the basic emotion responses are cross cultural, while responses to nonbasic emotions are learned, and hence culture dependent.

Three aspects can be determined as dimensions of the emotion ; "Strength" ranging from contempt to fear or surprise, "Valence" ranging from love or happiness to anger, and "Activity" ranging from sleep to tension.

Davitz (1964) found that the activity aspect of emotional meaning is carried by the relatively simpler elements of the vocal symbol, such as pitch and loudness, while both valence and strength are communicated by subtler and more complex vocal patterns of inflection, rhythm, etc. His study reported also that where erroneous judgments were made, it was very often in favor of another emotion with a similar activity level rather than a similarity in terms of valence or strength.

3. EXPERIMENTAL ANALYSIS

3.1. Procedure

The recordings were made in a recording studio using a high quality microphone and a DAT recorder. A male amateur actor and a male professional actor, about thirty years old, were recorded in separated session, each speaking the same sentences, five for each emotion, anger, joy, sorrow, or tenderness, and neutral as an unmarked expression. The actors were presented with a description of a situation causing a given emotion and asked to repeat the sentences six times. This procedure was same for the Korean and French recordings. The meaning of the sentences were : "Jean invited you at his party. Will you go there? You have met him last night, it is so often, isn't it? Don't go there, today. Let's prepare the dinner for my friends coming tonight."

In order to validate the expressivity of the recorded sentences and to know how the emotions are identified, we proceeded the tests of identification. Listeners were asked to identify the intended emotion among five ones, after listening a stimulus phrase ; the questionnaire consisted of thirty stimulus phrases presented once at five seconds interval. Four tests were carried out by the combination of two kind of corpus, Korean and French, and two groups of listeners, eighty-four Koreans and eighty-five French's.

In the result, the score of correct responses of the emotions was always in this order ; Anger > Sorrow > Tenderness > Neutral > Joy. The amateur expression of emotions and the professional were not significantly different in the listener's perception, even though the latter was more typical and obvious.

As to establish the Korean and French corpus to be studied, we selected ten sentences for each emotion with the high score identification.

3.2. Statistic Analysis

On account of limited space, we present mainly the result and the discussion of the imperative sentences of the corpus which were most representative to characterize each emotion ; the phonetic transcription of the Korean phrase is [onirin kaʒima] and the French [nivapa oʒurdyi].

By means of the analysis of variance (ANOVA) of the result, we found that the factors influencing the identification of emotions, such as emotion-type, comprehension level of a given language, and modality of the phrase, but the length of the phrase did not. The most identification was guaranteed when listeners heard a imperative phrase of anger in his mother tongue.

3.3. Acoustic Analysis

We studied three principal parameters, pitch , duration, and intensity, not missing the voice quality.

The pitch contour seemed to carry a large part of emotional information in both languages ; especially, anger was characterized by the abrupt final chute. The wide pitch range with dynamic movements of anger or joy was located higher than the narrow range of sorrow or tenderness. These differences were emphasized on the important words, for example, the negative morpheme of the imperative negative phrase, located at the final and the beginning of the Korean and the French phrase respectively.

The duration of the last syllable differed greatly according to the emotions : it was very short in anger and long in joy and tenderness, while the speech rate was high for anger and joy but slow for sorrow and tenderness.

Emotions	Active	Passive		Voice Quality
Positive	Joy	Tenderness	===>	resonant
Negative	Anger	Sorrow	===>	Aspirated / devoiced
	⇓	⇓		
	High	Low		Register
	Wide	Narrow		Pitch Range
	High	Low		Speech Rate
	High	Low		Intensity

Figure 1. Regrouping of the emotions according to their characteristics

The vowel lengthening occurred at the final part of every emotion : the last syllable was longest for tenderness and shortest for anger. However, the variation of consonantal duration was less regular ; usually consonants lengthened on important words.

The intensity was high for anger or joy and low for sorrow or tenderness, as expected.

The voice quality determined on the valance dimension was quite different depending on the positive or negative emotion : the joyful and the tender voice were more resonant than that of anger and sorrow which were more aspirated.

According to the characteristics common to the Korean and French sentences, we regrouped the studied emotions in active vs. passive emotions on the one hand, and positive vs. negative emotions on the other hand as described in the figure 1. Neutral sentences were closer to the passive emotion sentences than to the active. Similar experiences and results were reported by [1] and [2].

3.4. Perceptual Analysis

The previous tests of identification of emotions showed how the emotions are identified and confused. By reason of the complex paralinguistic features of the emotion, which are not coded systematically as linguistic features, it happens often the disjunction between speaker's coding and listener's decoding, interpreted as the confusion.

Pakosz (1983) noted five points of special interest from the literature :

"a) Speakers vary markedly in their ability to express emotive meaning vocally in controlled situations.

b) Listener's recognition and interpretation of emotions from recorded speech varies substantially.

c) Some emotions are more readily expressed and identified than others.

d) Misidentifications seem to follow a regular pattern whereby similarity on the activation dimension between two emotions leads to confusing one for the other.

e) Recognition of emotions is possible under conditions of reduced information concerning pitch variation."

These notes were revised and validated in our analysis. For instance, anger was quite well expressed in the imperative phrase and easily identified, even by foreigners, then less confused with other emotions. While joy was often confused, especially with anger having high intensity at high register in the activity dimension. In the same way, tenderness was confused with sorrow by their low intensity at the low register.

The previous regrouping is efficient to explain the fact of confusion : a given emotion is more often confused with emotions in the same group than in the other group, especially in the activity dimension. By the way, in view of the direction of confusion illustrated in the figure 2, it seems that listeners tend to choose a negative emotion when they are not sure to decide a positive emotion in the same group.

In the supplement test, we asked listeners to write adjectives evoked by the stimulus. Diverse adjectives were written for a given emotion and some of them were written again for the other emotions.

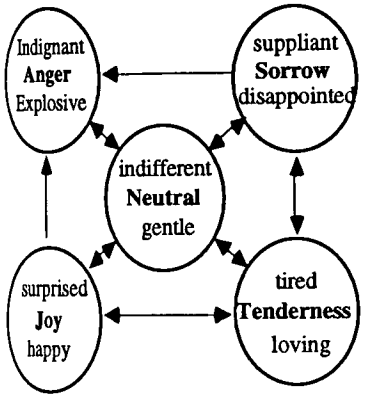

Figure 2. Direction of the confusion between emotions and Adjectives most frequently written in the supplement test

This figure shows the complex nature and the similarity of emotions. In a word, the perceptual confusion is caused by acoustic similarities of the emotions.

3.5. Filtering Effect

For the filtering experience, we selected the sentences of neutral and two emotions obviously contrasted, anger and tenderness, from the original recorded corpus.

In order to deprive the corpus of intelligibility, we eliminated the frequencies above 250 Hz. Apart from our intention, this operation affected the first formant of open vowels such as [a], then the syllabic regularity of intensity was disturbed. And it affected also some part of the F0 of anger exceeding 250 Hz, so destroyed its fine structure of pitch contour, while tenderness preserved its original pitch contour.

Next, we had listeners pass the test of identification of emotions with original and filtered corpus, of his mother tongue and of a foreign language, respectively ; thereby four tests were carried out.

As the result, the filtering had an effect on the identification of emotions when listeners heard the filtered corpus of his mother tongue. In most cases with unfiltered corpus, a listener identified better anger than tenderness, either of his mother tongue or of a foreign language. However, with the filtered, then unintelligible corpus, the pattern of the identification of emotions changed : the listener identified still better anger than tenderness in the filtered corpus of a foreign language but he identified better tenderness than anger in the filtered corpus of his mother tongue.

In conclusion, it seems that the fine structure of pitch contour and intensity play a important role in the identification of emotions and that listeners rely on different principal criterion, depending on the intelligibility of stimulus:

if the stimulus is intelligible with all information, whether he knows the language or not, he identified best anger having great intensity as well as fine pitch contour,

if the information is so reduced to make stimulus of a foreign language unintelligible, he identified still better anger than tenderness, based entirely on the striking intensity of anger

however, if he hears the unintelligible corpus of his mother tongue, his melodic intuition prevents him from the decision of anger which lost the fine prosodic structures of pitch and intensity, and he identifies better tenderness preserving the fine structure of pitch contour.

4. CONCLUSION

So far, we reported the analysis of the emotive speech in Korean and French, concerning the universality of emotions and the problem of perception.

This information could be used as the basis for a set of rules to control a high quality speech synthesizer with simulated emotion effects in the output speech. As emotion forms such an important part of human speech, its incorporation into speech synthesis systems is surely imminent.

REFERENCES

[1] Murray, I. R. and Arnott, J. L. (1993), *Toward the simulation of emotion in synthetic speech : A review of the literature on human vocal emotion*, J. Acoust. Soc. Am. 93 (2), pp. 1097-1108.

[2] Abadjieva, E., Murray, I. R., and Arnott, J. L. (1993), *Applying Analysis of Human Emotional Speech to Enhance Synthetic Speech*, Eurospeech 93, pp. 909-911.