Labeling expressive speech in L2 Italian: the role of prosody in auto-and external annotation

Marta Maffia, Elisa Pellegrino, Massimo Pettorino

Department of Literary, Linguistic and Comparative studies,
University of Naples "L'Orientale", Italy
{mmaffia, epellegrino, mpettorino}@unior.it

Abstract

The present study is intended to compare two approaches of labeling expressive corpora: auto-annotation and annotation by external lay listeners. These two methods have been applied to the semi-spontaneous emotional speech produced by Chinese learners of L2 Italian, involved in the CardTask, a mood-induction procedure that allows us to control the context of interaction, preserving the spontaneity of reactions.

The emotional responses to the stimuli presented in the task were the object of an auto-annotation session. The same samples were then administered only in the auditory mode to 20 Italian and 20 Chinese lay listeners. The results of perceptual tests have underlined some similarities and differences between both auto- and external annotation, and between the ratings given by external Italian and Chinese listeners. The labels chosen by native Italians were similar to those selected in the auto-annotation session, particularly in the case of anxiety, fear and disgust. The correspondence between the results of the two annotation methods may be ascribed to the different prosodic patterns characterizing the emotional states. The results of the annotation made by Chinese listeners show that they found it hard to give a specific emotional label to utterances produced in a second language relying solely on prosodic patterns.

Index Terms: L2 emotional speech, prosodic cues, emotion annotation methods

1. Introduction

The study of emotional speech poses many problems, both methodological and analytical. When dealing with the collection of expressive spoken corpora, the first issue to be addressed is the style of speech to analyze. According to Scherer taxonomy [1], three kinds of productions can be considered, each with some points of strength and weakness: recited, induced or spontaneous emotional speech. Other methodological problems concern the nature of the speakers involved (actors or naïf), the kind of stimulus for emotion elicitation (pictures, videos, dyadic interactions) and the linguistic materials analyzed (nonsense words, syllables, interjections, utterances) (see [2] for a review of the most common elicitation techniques).

In addition to these methodological difficulties, another issue to consider is the labeling of emotions, particularly in the case of authentic expressive speech. The most common approach is the annotation by human experts, trained to deduce labeling paradigm from theoretical hypotheses on the nature of emotions. The limits of this annotation technique, though presumed to be the most objective, have been already underlined [3]. Alternative methods used to label emotional speech are the administration of perceptual tests to naïve listeners and the auto-annotation technique, in which the speakers themselves are asked to judge their own emotional state. There is no doubt that the annotation method chosen has an impact on the labels obtained and that, in order to test the validity of emotional labels, various methods should always be combined [4].

Despite the complexities of this kind of research, the identification of acoustic correlates of emotions has been the object of many studies, both on a production and perception level (see [1], [5] for a review). A correlation between the activation dimension and the most frequently measured acoustic parameters has been demonstrated [6], [7]. High activation emotions (such as fear, joy, surprise and anger) are generally characterized by shorter pauses, a wider tonal range, higher values of F0 and intensity, and faster speech rate. Low activation emotions (for example sadness and disgust), by contrast, are vehiculated by longer pauses, a narrower tonal range, lower values of F0 and intensity, and a slower speech rate.

Moreover, cross-linguistic studies on emotional speech encoding and decoding have emphasized the role of prosodic features in the identification of different emotional categories and have indicated that specific emotional states are vehiculated by universal prosodic patterns [8], [9].

Although studies on vocal emotions have been conducted for a wide range of languages, L2 emotional speech has not yet been extensively analyzed, either from the acoustic or from the perceptual point of view. Previous researches on expressive interlanguage, carried out on learners of L2 English with different mother tongues, have focused mainly on the emotional force of swearwords, taboo words or love words, when pronounced or listened in a second language [10], [11], [12], [13]. The impact of emotional expressions in L1 and L2 has also been the object of psycho-physiological studies. Harris et al. [14], for example, monitored the skin conductance of Turkish–English bilinguals via fingertip electrodes while they were rating for pleasantness a variety of stimuli in Turkish (L1) and English (L2). The results of this study demonstrated a difference between L1 and L2 emotional forces, being more noticeable in late bilinguals [15], [16]. Harris' hypothesis, therefore, is that L1 is the language of emotional expressiveness, while L2 is that of emotional distance.

2. The study

The present study has a twofold objective: firstly it is intended to identify the acoustic features of expressive speech in L2 Italian on the basis of an auto-annotation labeling; secondly, in order to verify the effectiveness of the auto-annotation, the auto-labels are compared to those given by external Italian and Chinese listeners. To achieve this, 10 Italians and 10 Chinese learners of L2 Italian (C1 Level – CEFR [17]) were involved. They were all female, university students, aged between 18 and 23.

10.21437/SpeechProsody.2014-4

The high level of linguistic competence has allowed the non-native students' active participation in the task and, as a consequence, the collection of a large corpus of emotional speech in L2 Italian. The decision to recrute Chinese learners depended on the results of previous studies, according to which the expressive speech of Chinese speakers is characterized by a more moderate and restrained style than that of Italians [18].

In order to collect emotional speech, Italian and Chinese participants were involved in the CardTask, a mood-induction procedure that allows us to control the context of interaction preserving the spontaneity of reactions.

2.1. The CardTask

The CardTask is a speaking activity where a Giver and a Follower work in pairs. They sit at the same table but they cannot see each other because of the presence of a dividing panel. The Giver receives five cards; the Follower is given a deck of 25 cards. The Giver has to describe her five cards (fig. 1) and the Follower has two minutes to find each card in her deck. The Follower's task is rather difficult because the deck consists of very similar cards only differing from each other in small details (fig. 2).

Figure 1: *Giver's cards*.

Figure 2: *Examples of some cards in the Follower's deck*.

In the room there is also an experimenter, whose function is basically to ensure that the task is performed as planned and to control some disturbing unexpected events which occurred during the game [19]. The following events were designed to elicit emotional linguistic reactions in the players and to arouse five different emotions (anger, anxiety, disgust, fear and surprise):

1. at the beginning of the game, while the Follower is looking for the first card in the deck, the chronometer rings after only 20 seconds instead of the expected two minutes;
2. the experimenter pretends to find a big beetle (obviously fake) in the room;
3. the experimenter tries to leave the room but the exit door seems to be temporary locked;
4. the fifth card is not in the deck.

In this study, the CardTask game was organized in two sessions: one only involving Italian speakers, the other only Chinese participants. The interactions took place in the silent chamber of University of Naples "L'Orientale" and were videotaped.

2.2. Method

2.2.1 Acoustic analysis

In order to identify the emotional responses to the stimuli presented by the experimenter during the task, we extracted the utterances occurring in the time interval between the end of each stimulus and the resolution of the unexpected events. A total of 132 utterances of expressive speech was collected and, by means of Praat, the whole corpus was segmented and annotated in two tiers: syllables and speech runs.

For each speech portion, we measured the duration and number of syllables, the length of burst phenomena, silent and filled pauses, and the lowest and highest F0 values. We consider bursts as "very brief, discrete, non verbal expressions of affect in both face and voice triggered by clearly identifiable events" [20]. On the basis of these measures, we calculated the following indexes: articulation rate (AR) (syll/s), speech time composition (percentage of silence, disfluence, syllables and burst), and tonal range (st). Additionally, in order to highlight the F0 variations connected to the considered emotions, for every utterance we related the F0 min and max values to the lowest F0 value - the F0 floor (st) - reached by the speaker in the whole corpus.

In order to preserve the spontaneity of the interaction, we did not control the movements of the participants in the room and thus the production of background noises. For this reason we decided not to consider the intensity and voice quality features.

2.2.2 Labeling emotions

The extracted emotional speech samples, both in L1 and L2 Italian, were object of an auto-annotation session. Moreover, in order to obtain more reliable labels and to assess the perceptual effect of emotional speech in a second language, the auto-labels were compared to those given by external Italian and Chinese lay listeners (henceforth external annotation).

During the auto-annotation session, the Italian and Chinese participating in the CardTask were instructed to watch their video recordings and to annotate them with one of the six labels, the five expected emotions, plus a generic option "other". In the case of auto-annotation, the video was supposed to help the players contextualize the utterances and recall what they were feeling during the task.

As for the external annotation, since our attention was focused exclusively on the acoustic correlates of emotions in L2 and on their communicative effectiveness, the emotional utterances of Chinese speakers were administered only in the auditory form to 20 Italian and 20 Chinese listeners. They labeled the utterances following the same protocol adopted for the auto-annotation session.

In order to prevent Chinese participants from misunderstanding the relationship between the labels of emotions and referents of these labels, a native Chinese speaker, specialized in Italian language and linguistics, translated the emotion categories to their L1.

It is important to underline that the acoustic data presented in the following paragraph are organized on the basis of the auto-annotation labels.

3. Results of spectro-acoustic analysis

Figure 3 illustrates the mean values of tonal range and F0 height (register) for each emotion in L1 and L2 Italian.

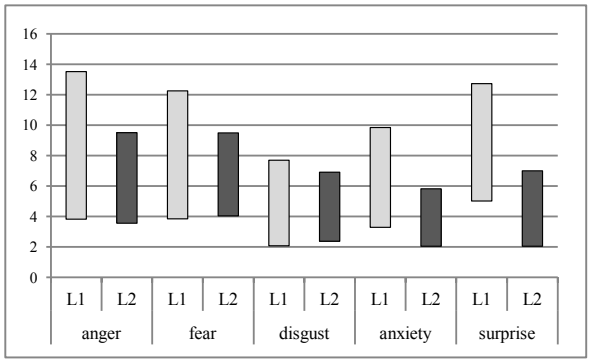

Figure 3: *Tonal Range and F0 height in L1 and L2 emotions (st).*

As the chart shows, the values regarding L1 Italian confirm the patterns for the high and low activation emotions, mentioned in the relevant literature. Of course anger, fear and surprise correlate with the highest F0 values and the widest tonal range. Anxiety, a high activation emotion as well, presents pitch height and tonal range values that are slightly lower than surprise, but higher than disgust, a low activation emotion [21].

Shifting our attention to the acoustic correlates of the emotions expressed by the Chinese subjects, it is possible to underline that the F0 values hardly match those attained by native speakers. As a matter of fact, F0 height and tonal range are quite steady in the whole corpus. The only exception is represented by anger and fear that are expressed with slightly higher values. These data seems to suggest that Chinese learners do not vary their pitch contour to distinguish different emotional states as in the case of native Italian speakers. After all, smaller pitch excursions are not only unique to L2 emotional speech, but they also represent one of the main acoustic correlates of Chinese accented Italian [22], [23].

Another parameter under study was articulation rate. Table 1 shows mean values for each emotional state in L1 and L2 Italian.

Table 1. *Articulation rate (syll/s) in L1 and L2 Italian.*

	Anger	Fear	Disgust	Anxiety	Surprise
L1 Italian	6.2	5.6	5.5	6.1	6.4
L2 Italian	5.2	5.6	4.3	4.8	4.8

Before considering AR variations in expressive speech, it is worth underlining that AR is a quite stable parameter, communicative setting being equal. With the exception of fear where L1 and L2 speakers reach the same values, in the other cases non-native utterances are produced with a slowing down of 1 or 1.5 syl/s, thus confirming the data available in the literature on foreign accented Italian [24]. The lower values of AR in L2 are essentially due to the learners' greater accuracy in uttering the single vowels and consonants. In order to reach the articulatory targets, the lengthening of syllable duration is needed, with a consequent slowing down of the speech.

Nevertheless, variations in AR do not seem to correlate with the five target emotions either in L1 or in L2. The only exception is represented by disgust in L2 Italian, whose values are particularly low (4.3 syll/s.) with respect to the whole corpus.

Figure 4 shows the composition of the utterance for each emotion in the two groups of participants.

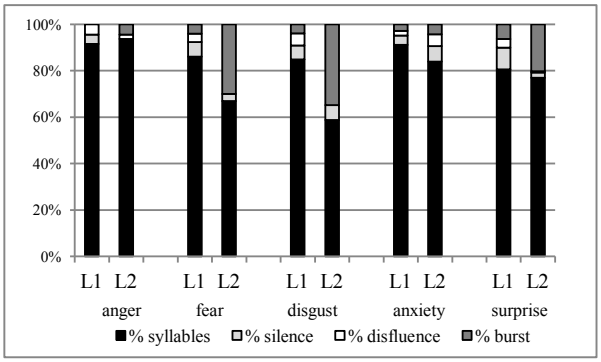

Figure 4: *Speech time composition in L1 and L2 Italian.*

In L1 Italian the percentage of syllables ranges from 85% to 95%, while in the second language it changes in the range of 60% and 93%. In the emotions that mostly differentiate L1 and L2 (fear, disgust and surprise), the bursts increase at the expense of syllable percentage. This means that when non-native speakers are not able to verbalize their emotions, bursts substitute the textual component. However, non-verbal expressions are not equally distributed between the five emotions. For example, in anger they are nearly absent, on the contrary, in disgust they represent one-third of the total time of the utterance. This datum is in line with Schröeder's findings [25]. Accordingly there are some emotions like disgust that are typically expressed by bursts, while others such as anger are not. A further observation is that in highly emotionally charged situations, L2 expressive speech is characterized by a very low percentage of disfluencies. L2 speakers rely on spontaneous and relatively universal vocal expressions to vehiculate their emotional states, instead of editing their performance with self-repairs, repetitions and substitutions. On the other hand, in L1, the percentage of disfluencies, above all in the form of lengthenings and vocalizations, is rather constant in the whole corpus. Such kind of filled pauses are typical of spontaneous speech and signal that the speaker is planning his/her speech.

4. Comparing auto- and external annotations

In order to verify the validity of labels given in the auto-annotation and to evaluate the communicative effectiveness of the selected expressive utterances in Chinese-accented Italian, we proceeded to compare the results of auto-annotation with the perceptual judgments given by external Italian and Chinese listeners. The results of the comparison underline some similarities and differences between auto- and external annotation and between the ratings given by native and non-native listeners. Tables 2 and 3 show the confusion matrices of auto- and external annotation by the two groups of listeners.

Table 2. *Confusion matrix of auto- and external annotation by Italian listeners*.

		External annotation				
		anger	fear	disgust	anxiety	surprise
Auto-annotation	anger	**29.2**	9.2	3.8	23.1	33.1
	fear	10.3	**48.3**	11.5	16.7	11.1
	disgust	5.8	25.0	**44.2**	7.7	16.3
	anxiety	6.1	14.2	2.4	**61.1**	13.4
	surprise	9.0	10.7	0.6	48.5	**28.6**

Table 3. *Confusion matrix of auto- and external annotation by Chinese listeners*.

		External annotation				
		anger	fear	disgust	anxiety	surprise
Auto-annotation	anger	**26.0**	8.0	4.0	28.0	34.0
	fear	5.6	**42.2**	8.9	8.9	34.4
	disgust	10.0	13.8	**30.0**	13.8	32.5
	anxiety	4.8	14.4	7.4	**66.7**	5.9
	surprise	9.4	9.4	6.1	47.2	**27.8**

As we can infer from the matrix of table 2, the labels chosen by the native Italians are pretty similar to those selected in the auto-annotation session by Chinese speakers, particularly in the case of anxiety, fear and disgust. The correspondence between the results of the two annotation methods may be ascribed to the specific prosodic patterns characterizing these emotional states. Anxiety is expressed by the lowest register and the narrowest tonal range. Fear and disgust are vehiculated by the same intonational patterns as in L1 Italian, though with not such marked differences. Moreover, disgust presents the lowest value of articulation rate and the widest portion of bursts. The similarities between values of F0 and articulation rate for surprise and anxiety induce listeners to confuse these two emotions. The emotional state that scores the highest percentage of mismatching between auto- and external annotations is anger, because of the considerable distance of its prosodic pattern from the model produced in L1 Italian. This result can be also explained by considering social conventions and different cultural standards in the expression of emotional states. As it has been noted in recent literature, Chinese speakers tend to inhibit the expression of emotions, which could threaten relational harmony as in the case of anger [8], [18].

The results of the annotation made by Chinese listeners show that they found it hard to give a specific emotional label to utterances produced in a second language. With the exception of anxiety, recognized by more than 60% of non-native listeners, the ratings given by Chinese subjects to the other emotions are definitely more uncertain than those expressed by Italians and more subject to random variations.

5. Conclusions

The present study had a twofold objective: firstly we intended to highlight the acoustic differences in the expression of the same emotions in L1 and L2 Italian; secondly we aimed to compare two approaches of labeling expressive corpora (auto- and external annotations) and verify the perceptual effects played by the prosodic patterns of L2 speech on native and non-native listeners.

As regards the acoustic correlates of emotional speech, the comparison between L1 and L2 confirmed Harris' hypothesis, according to which the second language is the language of emotional distance. Indeed the expressive speech of Chinese learners is characterized by a lower degree of variability in terms of F0 register and tonal range, a slowing down of articulation rate, and a different speech time composition. The reduced competence in the second language determines a lower percentage of syllable time and an increase of the burst component. During the CardTask, Chinese participants were not allowed to use their L1, the language of emotional expressiveness, but at the same time they were not able to use Italian to express their emotional states. Consequently, they tended to overcome this difficulty by relying on bursts.

As for the perceptual effectiveness of prosodic cues in the expression of emotions in L2 Italian, the data show that there is a correlation between the labels assigned by the speakers themselves in the auto-annotation and those chosen by native Italian listeners. This is particularly evident in the case of emotions characterized by similar prosodic patterns in L1 and L2 Italian (anxiety, fear and disgust). The lack of shared prosodic models, on the contrary, provokes in the listeners a high degree of uncertainty when labeling emotional states.

The random judgments given by Chinese external listeners reveal the objective difficulty of L2 learners to identify a specific emotional state in utterances produced in a second language.

Although intra- and cross-linguistic studies have already emphasized the difficulty of find coherent labels to the different emotions, this task seems to be much more demanding when analyzing emotional speech in a second language. The reason can be ascribed to the status of the L2 as the language of emotional distance, both on acoustic and perceptual levels.

In a further step of the research, we intend to extend the CardTask to learners with different mother tongues in order to verify whether the acoustic patterns characterizing the expression of emotions in L2 Italian are imputable to L1 transfer or are constrained by interlanguage development. The administration of perceptual tests to native and non-native speakers of Italian with different L1s will enable us to evaluate the extent to which the prosodic patterns and the cultural standards of the first language influence the perception of L2 emotional speech.

Further analysis of the corpus presented in this study will include a more detailed description of burst phenomena both in L1 and L2 Italian and the evaluation of the possible effect of the Chinese tonal system on the intonational features of the second language.

6. References

[1] Scherer, K. R., "Vocal Communication of Emotion: a Review of Research Paradigm", Speech Communication, 40:227-256, 2003.

[2] Coan, J. A., Allen, J. J. B. [Ed], The Handbook of Emotion Elicitation and Assessment, Oxford University Press, 2007.

[3] Aubergé, V., Audibert, N. and Rilliard, A., "Auto-annotation: an alternative method to label expressive corpora", in Proceedings of LREC, 2006.

[4] Truong, K. P., van Leeuwen D. A., Neerincx M. A. and de Jong F. M. G., "Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion", in Proceeding of: INTERSPEECH 2009, 10th Annual Conference of the

International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009.
[5] Johnstone, T. and Scherer, K. R., "Vocal communication of emotion", in M. Lewis and J. Haviland [Eds], Handbook of emotion (2nd ed.), 220-235, The Guilford Press, 2000.
[6] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijik, M. and Gielen, S., "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis", in Proceedings of the EUROSPEECH 2001, Aalborg, Denmark, 3-7 September, 1:87-90, 2001.
[7] Schröder, M., Speech and Emotion research. An overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis, PhD Dissertation, 2003.
[8] Yang, L., Campbell, N., "Linking form to meaning: the expression and recognition of emotions through prosody", in Proceedings of the 4th ISCA Workshop on Speech Synthesis, Perthshire, Scotland, 2001.
[9] Pfitzinger, H. R., Amir, N., Mixdorff, H., Bösel, J., "Cross-language perception of Hebrew and German authentic emotional speech", in Proceedings of ICPhS2011, Hong Kong, 1586-1589, 2011.
[10] Kaneko, T., "How non-native speakers express anger, surprise, anxiety and grief: a corpus-based comparative study" in M. Archer et al. [Eds], 384-393, 2003.
[11] Dawaele, J. M. and Pavlenko, A., "Emotion Vocabulary in Interlanguage", Language Learning, 52(2):263–322, 2002.
[12] Dawaele, J. M., "The Emotional Force of Swearwords and Taboo Words in the Speech of Multilinguals", Journal of Multilingual and Multicultural Development, 25(2-3): 204-222, 2003.
[13] Dawaele, J.M., "The emotional weight of I love you in multilinguals' languages", Journal of Pragmatics, 40(10): 1753-1780, 2008.
[14] Harris, C. L., Aycicegi, A. and Gleason, J. B., "Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language", Applied Psycholinguistics, 24: 561-578, 2003.
[15] Harris, C. L., "Bilingual Speakers in the Lab: Psychophysiological Measures of Emotional Reactivity", Journal of Multilingual and Multicultural Development, 2:223-247, 2004.
[16] Harris, C. L., Gleason, J. B. and Aycicegi, A., "When is a First Language More Emotional? Psychophysiological Evidence from Bilingual Speakers" in A. Pavlenko [Ed], Bilingual minds: Emotional experience, expression, and representation, 257-282, Clevedon, Mulilingual Matters, 2006.
[17] Council of Europe, The Common European Framework of Reference for Languages: Learning, teaching assessment, Cambridge University Press, 2001.
[18] Anolli, L., Wang, L., Mantovani, F. and De Toni A., "The Voice of Emotion in Chinese and Italian Young Adults", Journal of Cross-Cultural Psychology, 39:565-598, 2008.
[19] Maffia, M., Pellegrino, E., Vitale, M., De Meo, A., Pettorino, M.,"Expressive (Inter)language: A new method to elicit emotional speech in L1 and L2 Italian", paper presented at WASSS (Workshop on Affective Social Speech Signals), Grenoble 22-23 August 2013.
[20] Scherer, K. R., "Affect Bursts", in S. H. M. van Goozen, N. E. van de Poll and J. A. Sergeant [Eds], Emotions, 161-193, NJ: Lawrence Erlbaum, 1994.
[21] Jones, M., Anagnostou, F. and Verhoeven, J., "The vocal expression of emotion: an acoustic analysis of anxiety", in W.S. Lee, [Ed], Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, 17-21 August 2011, 982-985, 2011.
[22] De Meo, A. and Pettorino, M., "Prosodia e Italiano L2: cinesi, giapponesi e vietnamiti a confronto", in R. Bozzone Costa, L. Fumagalli and A. Valentini [Eds], Apprendere l'Italiano da lingue lontane: prospettiva linguistica, pragmatica, educativa, Guerra Edizioni, 59-72, 2011.
[23] De Meo, A. and Pettorino, M., "L'acquisizione della competenza prosodica in Italiano L2 da parte di studenti sinofoni", in E. Bonvino and S. Rastelli [Eds], La didattica dell'Italiano a studenti cinesi e il progetto Marco Polo, Pavia University Press, 67-78, 2011.
[24] Pellegrino, E., "The perception of foreign accent and speech. Segmental and suprasegmental features affecting degree of foreign accent in Italian L2", in H. Mello, M. Pettorino e T. Raso [Eds], Proceeding of the VIIth GSCP International Conference – Speech and Corpora, Firenze University Press, 261-267, 2012.
[25] Schröder, M., "Experimental study of affect bursts", Speech Communication, Special Issue following the ISCA Workshop on Speech and Emotion, 40:1-2, 99-116, 2003.