Automatic Analysis of Emotional Prosody in Mandarin Chinese: Applying the Momel Algorithm

Ting Wang [1,2], *Hongwei Ding*[1,3], *Qiuwu Ma*[1], *Daniel Hirst*[4,1]

[1] School of Foreign Languages, Tongji University, Shanghai, China
[2] Department of Linguistics, University of Pennsylvania, Philadelphia, USA
[3]School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China
[4]Laboratoire Parole et Langage, UMR 7309 CNRS, Aix-Marseille University, France
{2011ting_wang; hongwei.ding; mqw}@tongji.edu.cn; daniel.hirst@lpl-aix.fr

Abstract

Based on the Momel algorithm, a set of acoustic parameters was analyzed automatically on Chinese emotional speech. Global prosodic features were calculated on the sentence level, which showed a concordance with the usual pattern reported in the literature. Local constraints were also considered on the syllable layer. An ANOVA showed that there were interactive effects among emotions, syllable positions and syllable tones on certain parameters. Further more, by examining the pitch movements, no significant difference was found between neutral speech and active emotional speech, which was different from the performance in non-tonal languages. However when reducing the tonal influence by using utterances composed of only tone 1 syllables, this inverse effect disappeared. Hence we posited an interpretation that due to the existence of lexical tone in Mandarin Chinese, the paralinguistic use of pitch movements has been reduced.

Index Terms: emotional prosody; the Momel algorithm; Mandarin Chinese; lexical tones

1. Introduction

Human speech communication conveys not only linguistic information, but also shows the speaker's age, gender, emotions and other paralinguistic cues, among which the importance of emotions in vocal speech has been recognized throughout history. Emotion-specific patterns of acoustic cues have been widely investigated in non-tonal languages, such as in [1], [2] and [3], and showed some common properties across languages. However, the acoustic realization of emotion in lexical tone languages like Mandarin Chinese does not seem to always work the same way. [4] posited that the presence of lexical tones significantly constrains the manipulation of f_0 in emotional speech. Therefore, language-specific attributes, especially in tone language, still need more attention.

In Chinese, the lexical tones and intonation are intertwined as one phonetic representation of the raw f_0 curve conveying both linguistic and paralinguistic functions. How lexical tones and intonation interact with each other is still an unsolved problem. Chao [5] was one of the pioneers who studied Chinese emotional prosody. He pointed out that the emotional intonation depends on the voice quality, stress, phrase pitch and tempo of speech. Yuan et al. [6] proposed that anger and fear are mainly realized by phonation; joy is mainly realized by f_0; whereas sadness is realized by both phonation and f_0. Zhang et al. [7] investigated f_0, duration and short-time amplitude on both sentential level and syllable level. Li et al. [8] investigated emotional intonations by analyzing mono-syllabic utterances. Results showed that the tonal space, the edge tone and the duration differ greatly across 7 emotions. Wang et al. [9] studied the cross-linguistic perceptual patterns of four basic emotions. However, many questions are still waiting to be answered. For instance, the interaction between global intonation and local constraints like tone and position is still not quite clear. Furthermore, we should also find a proper way of modelling the emotional melody to account for microprosody and noise in f_0 extraction.

In this paper, the Momel algorithm was adopted to represent the surface f_0 contour in Chinese emotional speech. Model based acoustic analysis has been conducted on both global layer and syllable base. Finally, the restriction of lexical tone and vocal emotion realization was examined.

2. Method

The emotions used for investigation were four basic emotions including happiness, fear, anger and sadness, supplemented by a neutral state for comparison. Corresponding to the discrete emotion theory [10], these four emotions are among the most commonly postulated basic emotions and are most frequently studied [11]–[13].

2.1. Corpus

Recording materials are 36 sentences. Each sentence contains eight syllables. Target words differing in tones are put at the beginning, middle and end of each sentence respectively. All sentences are proper to elite different emotions under certain scenario. Examples of one group of 12 sentences are given below. The other two groups of 12 sentences have the same syntactic forms but with different contents.

{wang1/liu2/ma3/wei4} *ling2 ming2 wan3 xiang3 hui2 xue2 xiao4.*
"Wang/ Liu/ Ma/ Wei Ling wants to go back to school tomorrow evening."
zhe4 shi4 luo2 min3 de0 ***{xin1/ nan2/ nv3/ jiu4}*** *peng2 you3.*
"This is Luo Min's new/ boy/ girl/ old friend."
ma1 ma1 jiao4 xiao3 ming2 qu4 mai3 ***{mao1/yang2/niao3/lu4}****.*
"Mom asked Xiao Ming to buy a cat/ sheep/ bird/ deer."

The recordings were made in the sound booth at 48 kHz sampling rate with a 16-bit resolution. Two professional actors (one male one female) were recruited from Cinema College of Tongji University. The speakers were asked to act each utterance with each of the four emotional states and in a neutral way as a contrast. Each emotional state was accompanied by a short scenario with a picture. The elicitation scenarios help to minimize the interpretation variations that

10.21437/SpeechProsody.2014-12

may differ from speaker to speaker [14]. In total, we obtained 360 utterances (36 sentences×2 speakers×5 emotional states).

2.2. Listening test

Listening tests were conducted to confirm that the intended emotions were accurately decoded. Ten Chinese listeners were asked to choose for each the most suitable emotion among angry, happy, fear, sad and neutral options in a forced-choice task. Finally 303 utterances (72 for anger, 50 for fear, 56 for happiness, 53 for sadness and 72 for neutral) were chosen for further analysis.

2.3. Automatic alignment

For acoustic-related researches, one fundamental but time-consuming step before *real* analysis is the correct segmentation and alignment of the speech with the orthographic transcription. Several tools have been developed to make this laboursome task automatic, such as HTK [15], Julius [16] and the P2FA [17].

Here we chose a recently developed tool, SPPAS [18], to implement automatic phonetisation and alignment of Chinese emotional speech. SPPAS generated four tiers in the TextGrid including *inter-pausal units, words, syllables and phonemes*. For our Chinese speech, only three tiers were used in later analysis. An example is shown below:

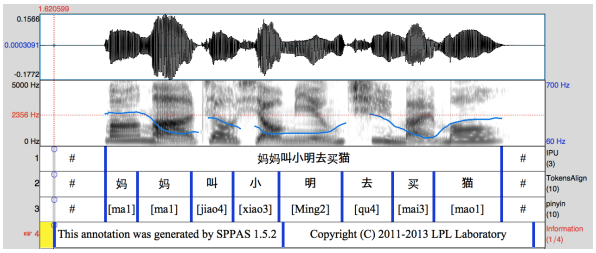

Figure 1: *SPPAS output example.*

2.4. The Momel algorithm

The Momel algorithm [19], [20], short for Modelling Melody, assumes that the raw f_0 curve is the interaction between two components: a global *macroprosodic* component determined by the underlying intonation pattern of the utterance, and a local *microprosodic* component caused by the articulatory constrains of segmental phonemes. Hence the underlying intonation pattern can be modeled as a continuous and smooth f_0 curve. In the Momel algorithm, this continuous curve is realized by a sequence of *target points* interpolated with a quadratic spline function.

The Momel algorithm has been applied to many languages including Standard Chinese, English, French, Korean, Italian, Catalan, Brazilian Portuguese, Venezuelan Spanish, Russian, Arabic and isiZulu [19]. In [21], this algorithm was first used on the lexical tone language, Standard Chinese, which showed its robust capability to model pitch contours.

The Momel algorithm is considered to be theory-neutral, or, better, *theory-friendly* [20], since it works as a phonetic representation of the intonation pattern with respect to speech production and speech perception. It can be compatible with some different theoretical approaches to describe speech prosody, and actually has been used as the first step in deriving the f_0 representations such as in the Fujisaki model [22], ToBI [23] and INTSINT [24].

Since this algorithm optimizes the modelling of speech prosody by taking the raw f_0 curves as input, and outputting the continuous and smooth macroprosodic components, it should be a powerful model to account for the rich local pitch changes in emotional speech.

2.5. Applying Momel algorithm to the corpus

The utterances in the emotional speech corpus were coded using the automatic Momel algorithm. Main steps were described in detail below.

2.5.1. Detect f_0

The quality of f_0 detection is a crucial fundamental step for all the pitch contour models. Unfortunately, most software like Praat will produce errors when extracting f_0 values if only the default parameters are used. Especially when there are a lot of pitch movements and for non-modal speech styles, f_0 extraction using the default parameters is not reliable. The most common errors are due to the inappropriate setting of minimum value and maximum value allowed for the f_0 analysis, that is, Pitch Floor and Pitch Ceiling in Praat.

The Momel plugin [19] on Praat provides an automatic f_0 detection algorithm to generate appropriate Pitch Floor and Pitch Ceiling parameters. This is a two-pass method. In the first pass, default parameters (50Hz for Pitch Floor, 700Hz for Pitch Ceiling) are used to calculate the f_0. Then, the first and third quartiles, named q1and q3, of the f_0 distribution are taken. In the second pass, Pitch Floor and Pitch Ceiling are recalculated respectively by the formula 0.75*q1 and 2.5*q3.

In this paper, we adopted this method by treating our emotional speech corpus with the automatic f_0 detection algorithm in Momel. The outputs were the corresponding *.Pitch* files. Instead of manual correction, this automatic way of detecting f_0 is preferable for the analysis of large speech corpus and is reproducible for further research.

2.5.2. Calculate Momel targets

Once we obtained the *.Pitch* files after the above step, a sequence of pitch target points was calculated by the Momel algorithm, which resulted in the *.PitchTier* files. Figure 2 shows an example of these pitch targets (black circles) detected in a happy utterance from our emotional speech corpus. The red curve is the modeled f_0 contour interpolated quadratically from the targets points with a quadratic spline function.

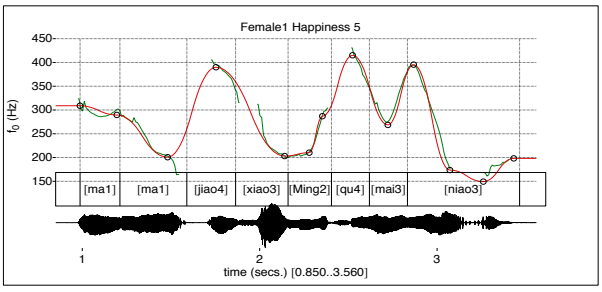

Figure 2: *Automatic Momel algorithm for the utterance "Mom asks Xiao Ming to buy a bird." Raw f_0 (green curves), pitch targets (black circles) and modeled f_0 (red curve).*

For best results, the output *.PitchTier* files containing pitch target points were manually checked using the *Correct Momel* interface. Any erroneous or missing target points were either deleted or added by hand.

3. Model-based acoustic analysis

Melody metrics [19], [20] were extracted by a Praat script based on the Momel outputs on the unit of sentence or syllable, including parameters of duration, intensity, f0 and their variance measures which are regarded as the basic parameters of speech prosody.

3.1. Global prosodic pattern

We first used sentence level as the analysis unit to capture the global prosodic pattern under each emotional state. The averaged results are listed in table 1.

Table 1: *Summary of the acoustic measurements*

	Anger	Fear	Happiness	Sadness	Neutral
Intensity (dB)	68.64	65.54	67.86	63.81	65.00
Speech rate (syl/s)	4.904	4.270	4.105	3.046	3.469
f_0 (octave in z-score)	-.043	.974	.784	-.238	-1.067
f_0 range (octave in z-score)	.261	-.481	.752	-.594	-.075
Mean abs slope (octave/s)	2.447	1.607	2.30	1.500	2.470

A one-way ANOVA with emotions (5 levels) as factor was conducted on the parameters described above.

For intensity, there was a significant main effect of emotions, F (4, 298) = 33.897, p = 0.000<0.05. The mean intensity of anger and happiness were significantly higher than others, while sadness is significantly lower, as indicated by post hoc multiple comparisons. For speech rate, there also appeared main effect of emotions, F (4,298) = 68.730, p = 0.000 < 0.05. The utterances spoken in an angry way were significantly faster, followed by fear and happiness. The speech rate under the sadness state was the lowest, and we only found inner-sentence pauses in sad speech. f_0 and f_0 range values were calculated on interpolated *.PitchTier* files from Momel. All values were first converted from Hz to octave, and then z-transformed with each speaker to reduce the inter-subject variability. A main effect of emotions was found for both parameters, F (4, 298) = 92.831, p = 0.000 < 0.05 and F (4, 298) = 21.436, p = 0.000 < 0.05 respectively. Post hoc tests showed that happiness and fear had significantly higher mean f_0, followed by anger, sadness and neutral. Anger had largest f_0 range, then happiness larger than neutral significantly. Fear and sadness had no significant difference.

Mean absolute slope was measured as absolute difference from the previous pitch point divided the by distance in seconds, which indicated whether there were a lot of pitch movements or not. A significant main effect was shown across different emotions, F (4, 298) = 21.381, p = 0.000 <0.05. In post hoc tests, we found that neutral, anger and happiness had significantly more pitch movements than fear and sadness as illustrated in Figure 3. What drove us to pay attention was that the mean absolute slope of neutral versus anger and happiness in Chinese showed a different pattern to that found in other non-tonal languages in [25], [26]. As reported in the literature, active emotional speech, such as anger and happiness, have much more pitch movements than neutral speech. However, we didn't find such an effect in our Chinese corpus. Either because speakers in our study used a different strategy to express these emotions, or because there's some underlying difference of pitch movements across emotions between Chinese and other non-tonal languages. Detailed discussion will be given later.

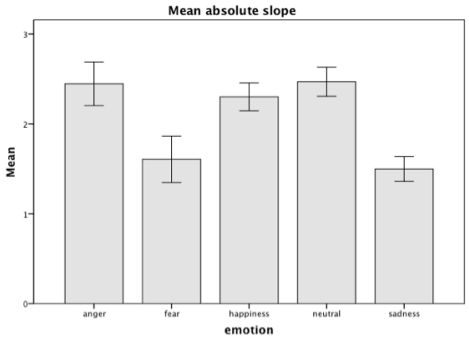

Figure 3: *mean absolute slope on sentence level.*

In summary, as for the sentence level, different emotional speech showed significantly different acoustic patterns, which are largely in concordance with the usual patterns reported in the literature, such as in [2] and [27]. However, the amplitude of pitch movements indicated by mean absolute slope showed different results, which required further analysis.

3.2. Emotions, positions and tones interaction

Although many of the acoustic studies focus on the global realization of prosody, studies on synthesis of emotional speech didn't work like this. A bottom-up method has been adopted. For synthesis purpose, different parameters were usually manipulated on the levels of syllables or even phonemes. Here we should draw attention to the local constraints. In this study, different tones and sentence positions were considered together with emotions on the acoustic analysis of target words/syllables in the utterances.

Table 2 shows the significant level for each of the acoustic parameters analyzed by ANOVA, with emotions (5 levels), positions (3 levels: beginning, middle and end of the sentence) and tones (4 levels: tone 1, tone 2, tone 3 and tone 4 in Mandarin Chinese) as factors. There were significant main effects of emotion, position and tone on most of the parameters except for mean slope. Significant effects of interaction between factors were also found on some parameters. Due to the limitations of space, detailed analysis of these results will be postponed for future publication, instead we focus on f_0 and mean absolute slope.

For mean f_0, the interaction between emotions and positions, emotions and tones were significant, F (8, 243) = 7.985, p = 0.000 < 0.05 and F (12, 243) = 2.519, p = 0.004 < 0.05 respectively. Figure 4 shows the mean f_0 performance in z-scored octave at different tones and emotions.

Across all emotions, tone 1 and tone 4 had higher f_0, which is in line with the fact that in Mandarin Chinese tone 1 and tone 4 begin as high tones. Interestingly, if look at the amplitude difference between tone 1, tone 4 and tone 2, tone 3, we found that compared with neutral speech, this difference is reduced in emotional speech. This phenomenon suggested the possibility that emotions, to some extent, restricted the distinction between the four tones in Chinese. This result seemed to provide an explanation for the unusual performance

of pitch movements described in 3.1. When speaking with emotions, pitch movements were reduced compared to neutral speech in Mandarin Chinese. To further investigate this phenomenon, we did an additional experiment.

Figure 4: *mean f_0 value (z-scored octave) at different tones and emotions.*

3.3. Lexical tone and vocal emotion restriction

To test the hypothesis above, we recorded four groups, each for one tone, of monotone utterances in Mandarin Chinese using the same female speaker. The utterances are digital strings, which also contain eight syllables for each. The recording procedure was the same as the previous experiment, which yielded 189 utterances. One example is as follow:

yi1 ba1 ba1 san1 qi1 yi1 ba1 qi1.

"One eight eight three seven one eight seven."

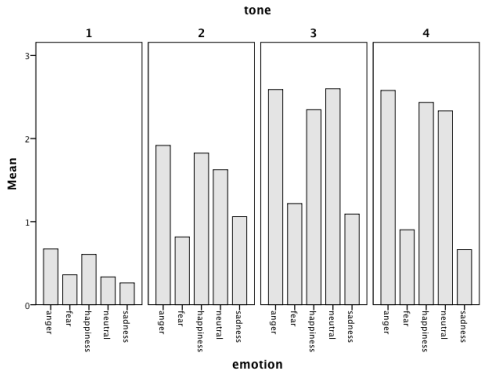

Figure 5: *mean absolute slope in monotone utterances.*

Mean absolute slope was calculated on Momel output as illustrate in Figure 5. In utterances with only tone 2 or tone 3 or tone 4, we found the same pattern as in Figure 4 that, neutral, anger and happiness had significantly more pitch movements than fear and sadness. However, in utterances with tone 1, anger and happiness had significantly more pitch movements than neutral, which was consistent with the pattern in non-lexical tone languages [25], [26]. Since tone 1 in Mandarin Chinese is a level tone (usually represented as 55), the utterances with only tone 1 syllables, to some extent, reduce the tonal influence on intonation compared to other tone combinations, and can be regarded as similar to a non-tonal utterance. From the evidence above, we came to the conclusion that the lexical tone in Mandarin Chinese has an influence on the realization of emotional prosody. It seems that the lexical tone restricts the paralinguistic use of pitch. This point has also been reported from other evidences in [28]. A similar result was also shown for Cantonese [29] which has a richer tone system than Standard Chinese.

4. Discussion and conclusion

Based on the Momel algorithm, a set of acoustic parameters has been analyzed automatically on Chinese emotional speech. Global prosodic features were calculated on the sentence level, which showed a consistency with the usual pattern reported in the literature. Given the local constraints, we further examined the acoustic performance on the syllable level. An ANOVA showed that there were interactive effects among emotions, syllable positions and syllables tone on certain parameters.

Although previous findings such as [30] showed evidence for the existence of universal patterns from vocal characteristics to specific emotions across cultures, the existence of language-specific paralinguistic features were still found in vocal emotion expression. By examining the pitch movements indicated by mean absolute slope, no significant difference was found between neutral speech and active emotional speech, which was quite different from the performance in non-tonal languages. However when reducing the tonal influence by using utterances composed of only tone 1 syllables, this inverse effect disappeared. Hence we posited an interpretation that due to the existence of lexical tone in Mandarin Chinese, the paralinguistic use of pitch movements has been reduced. This result served as a further proof of the finding in [28] and [29].

The automatic phonetic representation of the intonation pattern, with respect to both speech production and speech perception, by the Momel algorithm makes it possible to account for more tonal and non-tonal language comparison in the future. More over, we should take into consideration more local constrains such as narrow focus and tone sandhi.

5. Acknowledgements

The first author benefited from the Scholarship Award for Excellent Doctoral Student granted by Ministry of Education of China, and the scholarship from the China Scholarship Council. This research was also supported by the National Social Science Foundation of China (No.13BYY009) and Innovation Program of Shanghai Municipal Education Commission (No. 12ZS030).

Table 2: *Significance levels of ANOVA for each parameter. [--]:no significance,[*]=p¡0.05,[**]=p¡0.01, [***]= p¡0.001*

	emotion	position	tone	emotion*position	emotion*tone	position*tone	emotion*position*tone
Intensity	***	**	***	***	--	--	--
Duration	***	**	***	***	--	***	--
f_0	***	***	***	***	**	--	--
f_0 range	*	***	***	--	*	*	*
Mean slope	--	--	***	*	--	***	--
Mean abs slope	***	***	***	--	--	--	--

6. References

[1] Murray, I. R., & Arnott, J. L., "Toward the simulation of emotion in synthetic speech : A review emotion", *The Journal of the Acoustical Society of America*, 93, pp. 1097–1108, 1993.

[2] Johnstone, T., & Scherer, K. R., "Vocal communication of emotion", *Handbook of emotions*, 2000.

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

[4] Ross, E. D., Edmondson, J. A., & Seibert, G. B., "The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice," *Journal of Phonetics*, 14(2), 283-302 1986.

[5] Chao, Y. R., "Tone and intonation in Chinese", *Bulletin of the Institute of History and Philology*, 4(2), 121-134, 1933.

[6] Yuan, J., Shen, L., & Chen, F., "The acoustic realization of anger, fear, joy and sadness in Chinese", in *INTERSPEECH*, pp. 2025–2028, 2002.

[7] Zhang, S., Ching, P. C., & Kong, F., "Acoustic analysis of emotional speech in Mandarin Chinese", *In International Symposium on Chinese Spoken Language Processing, pp. 57-66*, 2006.

[8] Li, A., Fang, Q., & Dang, J., "Emotional intonation in a tone language: Experimental evidence from Chinese", *ICPhS XVII, Hong Kong*, 2011.

[9] Wang, T., Ding, H., & Gu, W., "Perceptual Study for Emotional Speech of Mandarin Chinese", *in Speech Prosody 2012*, 2012.

[10] Darwin, C., "The expression of the emotions in man and animals". Oxford University Press, 1998.

[11] Ekman, P., "An argument for basic emotions", *Cognition & Emotion*, 6(3-4), 169-200, 1992.

[12] Ekman, P., "Are there basic emotions?" *Psychological Review*, 99(3), 550-553, 1992.

[13] Ekman, P., "Basic emotions", *Handbook of cognition and emotion*, 4, 5-60, 1999.

[14] Pell, M. D., "Influence of emotion and focus location on prosody in matched statements and questions", *The Journal of the Acoustical Society of America*, vol. 109, no. 4, p.1668, 2001.

[15] Young, S. J., & Young, S., "The htk hidden markov model toolkit: Design and philosophy", *Entropic Cambridge Research Laboratory, Ltd*, 1994.

[16] Lee, A., Kawahara, T., & Shikano, K., "Julius---an open source real-time large vocabulary recognition engine", *in EUROSPEECH* 2001, 2001.

[17] Yuan, J., & Liberman, M., "Speaker identification on the SCOTUS corpus", *Journal of the Acoustical Society of America, 123(5), 3878*, 2008.

[18] Bigi, B., & Hirst, D. J., "SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody", *Proc. of Speech Prosody*, 2012.

[19] Hirst, D. J., "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation", *In Proceedings of the XVIth International Conference of Phonetic Sciences*, pp. 1233–1236, 2007.

[20] Hirst, D. J., "The analysis by synthesis of speech melody: from data to models", *Journal of speech Sciences*, vol. 1, no. 1, pp. 55–83, 2011.

[21] Zhi, N., Hirst, D. J., & Bertinetto, P. M., "Automatic analysis of the intonation of a tone language. Applying the Momel algorithm to spontaneous Standard Chinese (Beijing)", in *INTERSPEECH* 2010, 2010.

[22] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", *In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference, Vol. 3, pp. 1281-1284,* 2000.

[23] Maghbouleh, A., "ToBI Accent Type Recognition", *ISSUES*, 1998.

[24] Hirst, D. J., "La représentation linguistique des systèmes prosodiques: une approche cognitive", *Doctoral dissertation, Aix Marseille 1,* 1987.

[25] Paeschke, A., Kienast, M., & Sendlmeier, W. F., "F0-contours in emotional speech", *Proc. ICPhS*, 1999.

[26] Paeschke, A., & Sendlmeier, W. F., "Prosodic Characteristics of Emotional Speech : Measurements of Fundamental Frequency Movements", *In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[27] Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P., "Emotional speech: Towards a new generation of databases", *Speech communication,* 40(1), 33-60, 2003.

[28] Hirst, D. J., "Melody metrics for prosodic typology: comparing English , French and Chinese", *INTERSPEECH 2013*, 2013.

[29] Hirst, D. J., Wakefield, J. & Li, H.T.Y. "Does lexical tone restrict the paralinguistic use of pitch? Comparing melody metrics for French, English, Mandarin and Cantonese". in *Proceedings of the International Conference on the Phonetics of the Languages in China*, Hong Kong, 2013.

[30] Scherer, K. R., Banse, R., & Wallbott, H. G., "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural psychology, 32(1), 76-92*, 2001.