applied sciences

MDPI

Article

Cross-Cultural Perception of Spanish Synthetic Expressive Voices Among Asians

Ganapreeta Renunathan Naidu [1], Syaheerah Lebai Lutfi [1,*], Amal Abdulrahman Azazi [1], Jaime Lorenzo-Trueba [2] and Juan Manuel Montero Martinez [2]

[1] School of Computer Sciences, University Sains Malaysia, Gelugor 11800, Pulau Pinang, Malaysia; grn14_com028@student.usm.my (G.R.N.); aaaa11_com017@student.usm.my (A.A.A.)
[2] Speech Technology Group, ETSI Telecomunicacion, Universidad Politécnica de Madrid, Calle Ramiro de Maeztu, 7, 28040 Madrid, Spain; jaime.lorenzo@die.upm.es (J.L.T.); juancho@die.upm.es (J.M.M.M.)
* Correspondence: syaheerah@usm.my; Tel.: +60-4653-4388

Received: 7 November 2017; Accepted: 25 December 2017; Published: 12 March 2018

Abstract: Nonverbal cues play a vital role in contributing to how emotions are perceived, especially by outgroups. In this study, a cross-cultural perception experiment of Spanish Synthetic Expressive Voices (SEV) was conducted to investigate the perception rate among different groups of Asians towards the SEV. Ten (10) subjects from each ethnic group namely Japanese, Chinese, Vietnamese, and Malaysians participated in this test. The subjects were required to listen to and categorize the SEV corpus which contains 260 utterances with 4 emotions (*anger*, *happiness*, *sadness*, and *surprise*) and the *neutral* speech in different intensities and durations. Overall, the results indicate that duration and intensity of speech plays a significant role in perception. This paper concludes that listeners' perceptions are influenced by a speaker's nonverbal expression and it is important that these features (duration and intensity of speech) are considered when modelling synthetic speech for artificial agents in real-time applications in a cross-cultural user environment.

Keywords: nonverbal; Spanish; perception; prosody; cross-cultural effects; artificial agents

1. Introduction

Nonverbal communication is defined as interaction without language. Thus, vocal sounds that are not words, such as grunts or singing a wordless note, are nonverbal [1]. The human voice contains a rich source of nonverbal vocal expressive cues which come in different forms such as tone (prosodic expressions), breathing sounds, crying, hums, grunts, laughter, shrieks, and sighs (e.g., paralinguistic cues) [2,3]. Extensive reviews have established that the prosodic expressions of basic emotions—such as anger, fear, happiness, and sadness—are elicited by the acoustic patterns of cues related to pitch, intensity, voice quality, durations, and speed rates [4,5].

Listeners are able to interpret the emotional content of prosodic variations across languages and cultural differences with accuracy above chance [5–9]. Studies on speech particularly suggest that the perception of prosodic expressions has a universal component, although further meta-analyses have attested that judges rated expressions from people of their own culture (in-group) more accurately than they did with unfamiliar cultures [4,10]. Interactions with in-group members (people from the same culture) are characterized by a sense of intimacy, familiarity, and trust while interactions with out-group members (people from a different culture) lack these qualities.

The current study focuses on listener's perception of vocal emotion expressed in synthesized voices based on nonverbal vocal cues. Nonverbal cues play an important role for speakers and listeners to express and recognize emotions. To achieve this goal, we conducted a case study using the Spanish Synthetic Expressive Voices database (SEV) [11] to measure the emotion perception accuracy among

the following Asian cultures: Japanese, Chinese, Vietnamese and Malaysians. The effect of two of the prosodic features (intensity and duration) on the perception accuracies was further studied to find the common patterns of emotion perceptions among Asians.

Related Work

In some situations, unintentionally speakers tend to communicate emotions through many nonverbal cues which derive mainly from facial and vocal behaviors [12]. In a conversation, vocal cues of a speaker dramatically influence the perception and appraisal of a social interaction especially when it involves out-group listeners. Joy is expressed with a comparatively rapid speaking rate, high average pitch, and large pitch range. Sadness is conveyed with a slow speaking rate, low average pitch, narrow pitch range, and low intensity. Anger is expressed with a fast speaking rate, high average pitch, wide pitch range, high intensity, and rising pitch contours. Fear is conveyed with a fast speaking rate, high average pitch, large pitch variability, and varied loudness [13]. These emotional cues are well documented for English speakers; however, there might be certain differences in vocal qualities that are associated with emotion in other languages (e.g., Asian languages).

Prosody features (intensity, vocal pitch, rhythm, rate of utterance) contribute in forming the different expressions of emotions which are used to decode the meaning of a person's speech [14–16]. Elfenbein and Ambady (2002) found that members who shared similar cultural elements (e.g., degree of individualism or collectivism, power structure, and gender roles) are more successful at interpreting each other's emotional expressions than non-members of that culture. For example, cultural closeness predicts that Japanese people should be better at recognizing the emotional expressions of Chinese people than the emotional expressions of Americans because Japanese and Chinese cultures are more similar in terms of relevant dimensions compared to Japanese and American cultures [17]. This discrepancy indicates that socio-cultural influences on emotion recognition are noticeable in the vocal channel due to the unique interplay of emotion and language in speech. In this context, "natural" cues to emotion in the voice are tightly intertwined with the acoustic-phonetic properties of individual speech sounds [9] In a similar vein, linguistically assigned differences in the segmental inventory and intonational features, which contain the accent and rhythmic structure of a foreign language, could also impact emotion recognition by interacting negatively with the basic aspects of auditory speech processing or with the processes for extracting salient emotional features from prosody [9].

In a perception study by Scherer et al. (2003), native speakers from nine different countries across three regions (e.g., Europeans and Malays-Indonesians) were presented with 30 emotionally-modulated but semantically-anomalous utterances produced by four German actors. The researchers found that all listener groups recognized fear, joy, sadness, anger, and neutral utterances strictly from prosody at above-chance accuracy. The results emphasized that German (native) listeners performed significantly better on this task compared to the other language groups. It was apparent that language similarity appeared to influence vocal emotion recognition because listeners whose native language was more linguistically similar to German (e.g., Dutch) tended to be more accurate than those from a highly dissimilar language (e.g., Malay). These data imply that language and cultural features play an important role in how vocal emotions are recognized and decoded [10].

Another experiment by Beier and Zautra (1972) investigated emotion perception using English sentences that were of different length and accent [18]. The study revealed that in-group listeners (Americans) outperformed out-group listeners (Japanese) beyond chance level when listening to shorter utterances (e.g., "hello, how are you?"). However, this in-group advantage was not apparent when recognizing emotions from longer sentences, suggesting that the duration of an utterance affects the perception rate. In a similar vein, evidence from a study investigating speech rate demonstrated that the articulation rate of northern American speech was significantly faster than that of southern speech, both in reading and speaking [19]. Faster speaking rates show greater difficulty in processing

spoken words than slower rates do [20]. Slow speech rate influences the accuracy rate of translating information because it allows the listener more time to process the message [21].

Many researchers have also acknowledged that the experiences differ. "Listening to someone speaking clearly on a subject you know in a language you understand is a quite different than hearing someone say the same thing in a language you do not understand", McCulloch, 1993 (p. 46 [22]). In-group and out-group speakers of a language hold presumptions with regard to foreign speech—a notion perhaps due to the low intelligibility or complexity of accented speech by native speakers. Such biases may even apply to native speakers with a dialect different from the standard language, thus triggering misperception and discrimination [23]. Regional dialect can be a strong predictor of between-speaker variation in emotion recognition rate. A language spoken in a particular accent with specific intonation or timing properties thus contains important cues that need to be considered when modelling synthetic speech that contributes to the perception of emotion, especially to out-group listeners [9]. Language and individuality are not independent. The way of speaking (including how one pronounces particular sounds, how one conjugates verbs, speaking in a sing-song voice, and the formality of one's language) and the use of nonverbal behavior (e.g., gestures, use of the eyes and hands) differ from one culture to another [24,25].

With the advancement of technology, the use of artificial agents in various fields such as learning, banking and entertainment has become a norm [26]. Hence, these agents are expected to be equipped with the ability to communicate and respond with users based on their social and cultural background [27]. The voice of an agent as well as its embodiment has a strong impact on human perception [28]. Thus, artificial agents capable of natural language interactions are essential for establishing a trustable and comfortable environment for users of different ethnic backgrounds.

The work presented in this paper contributes to the view that human-like agents enhance the processes of social interpretation and evaluation, whereby the use of natural language—in particular speech variability and cultural features—are important factors that should be considered when modelling agents in a cross-cultural setting.

2. Experiment

A listening test was conducted to study the perception of emotional states using the Spanish Synthetic Expressive Voices (SEV) database [29]. This initial experiment was conducted with 55 evaluators who were native Spanish speakers aged between 15–65 years old. The setting of the experiment and evaluation procedure for this initial experiment may have varied from the current experimental method.

2.1. The Corpus

The Spanish Synthetic Expressive Voices (SEV) is a multimedia and multi-purpose database designed for research on emotional speech. The SEV used for this experiment is an acted expressive speech compendium using multi-emotion speakers which was produced and recorded by Grupo de Technología del Habla (GTH)—Speech Technology Group, University Politécnica de Madrid, Spain [30].

The SEV is a Spanish emotional speech corpus consisting of both a male and a female speaker recorded in several emotional states, including neutral speech. All the texts used for this corpus were devoid of emotional content. Out of all the available data, only the male speaker's voice in four emotions (anger, happiness, sadness, and surprise) and the neutral speech were considered for this experiment.

This speech corpus used was recorded using a professional male voice actor to produce utterances that contain 20 Spanish sentences in 3 intensities (0.5, 1.00, 1.25 decibels) for each category of emotions. Each sentence has 1 utterance in neutral state and 3 utterances (duration: 0.02–0.06 s) in 4 emotions (anger, happiness, sadness, and surprise). Therefore, in total, 260 sentences were used in this experiment.

The recording of the SEV was done in high quality acquisition set-up in a recording acoustic treated chamber. The equipment used was able to record 20 audio channels and a video signal at the same time. All audio, video recordings in this speech corpus (SEV) were fine-tuned using computer applications to establish quality speech with natural emotional content. The SEV has also been labelled (phonetically) automatically using suitable software [30].

2.2. Subjects

This experiment was conducted with a total of 40 subjects from 4 different ethnic groups: Japanese, Chinese, Vietnamese and Malaysians (10 subjects per ethnic group). The subjects were University students aged between 20–30 years old. All the subjects did not have exposure to or understand the Spanish language.

2.3. Method

In this experiment, only voice modality was used to investigate the perception rate among different ethnic groups. The subjects were asked to listen to the expressive utterances (SEV) and then identify the perceived emotions. The listening test was conducted in a sound-proof audio room to eliminate noise. The experiments with Japanese, Chinese, and Vietnamese subjects were conducted at the audio room at the Unoki & Akagi Laboratory, Japan Advanced Institute of Science and Technology (JAIST), whereas the experiment with the Malaysian subjects took place at the audio room at the School of Computer Sciences, Universiti Sains Malaysia. Both experiments were conducted under similar environments and settings.

The subjects were asked to use a headphone to listen to the audio files. The total time taken by each subject for this experiment was approximately 45 min. The subjects were let to proceed in the experiment at their own pace and were allowed to repeat the same audio file as many times as desired before recording their results and proceeding to the next audio file. Every subject was exposed to the same environment when the listening test was conducted, as shown in Figure 1.

Figure 1. Example of the listening test environment.

3. Results and Discussion

The data collected from the listening test were analyzed statistically in SPSS [31]. The emotion perception accuracies were calculated using the F1-measure because the numbers of records were not equal in each emotion class. The differences between variables—ethnic, duration, and intensity—were analyzed with the univariate analysis of variance (ANOVA) while the multivariate analysis of variance

(MANOVA) was applied to explore the interaction between the independent variables obtained from the experiment conducted for the "Simple4All" project [29]. These results have not been included in the ANOVA or MANOVA analysis for assessment because the experiment was conducted in the earlier phase and did not include assessment of more fine grain features such as intensities and duration.

3.1. Overall Results on Emotion Perception

Table 1 illustrates the confusion matrices of emotion perception by the five ethnic groups: Spanish, Japanese, Chinese, Vietnamese, and Malaysian. The following subsections discuss the results according to the ethnic groups. A graphical comparison in percentages between the actual emotion perception accuracy by the five ethic groups is illustrated in Figure 2.

Table 1. Confusion Matrix—Actual Emotion Against Perceived Emotion by Spanish, Japanese, Chinese, Vietnamese and Malaysian Subjects (based on the listening test using the SEV Corpus).

		Anger	*Happy*	*Neutral*	*Sad*	*Surprise*
Anger	**Spanish**	*** 71.20%**	0.00%	7.30%	3.60%	7.30%
	Japanese	**22.00%**	2.50%	58.70%	12.00%	4.80%
	Chinese	**43.20%**	5.50%	35.80%	10.50%	5.00%
	Vietnamese	**27.50%**	1.80%	49.80%	15.50%	5.30%
	Malaysian	**48.00%**	9.20%	29.80%	4.80%	8.20%
Happy	**Spanish**	10.90%	**32.70%**	10.90%	12.70%	25.50%
	Japanese	25.30%	**33.00%**	20.70%	4.00%	17.00%
	Chinese	26.30%	*** 38.20%**	18.20%	4.00%	13.30%
	Vietnamese	27.30%	**35.70%**	20.20%	1.70%	15.20%
	Malaysian	31.30%	**34.00%**	18.00%	3.70%	13.00%
Neutral	**Spanish**	7.30%	1.80%	**76.30%**	1.80%	14.60%
	Japanese	9.50%	1.50%	**73.50%**	14.50%	1.00%
	Chinese	4.50%	3.00%	**77.00%**	11.00%	4.50%
	Vietnamese	2.50%	0.50%	*** 84.00%**	9.00%	4.00%
	Malaysian	7.50%	5.00%	**75.00%**	7.50%	5.00%
Sad	**Spanish**	0.00%	0.00%	16.40%	*** 78.10%**	3.60%
	Japanese	3.30%	2.20%	21.30%	**71.00%**	2.20%
	Chinese	2.20%	3.20%	19.30%	**72.70%**	2.70%
	Vietnamese	1.50%	0.80%	27.80%	**67.70%**	2.20%
	Malaysian	3.30%	3.50%	20.30%	**70.20%**	2.70%
Surprise	**Spanish**	14.60%	30.90%	9.10%	3.60%	**38.10%**
	Japanese	11.00%	24.20%	28.30%	1.70%	**34.80%**
	Chinese	13.20%	29.50%	21.30%	2.80%	**33.20%**
	Vietnamese	11.20%	20.20%	25.20%	1.50%	**42.00%**
	Malaysian	11.00%	15.50%	23.80%	3.80%	*** 45.80%**

* Denotes the highest percentage of recognition rate among ethnic groups.

3.1.1. In-Group Perception Results (Spanish)

Certainly, the perception results by Spanish subjects (refer to Table 1) should be higher than other ethnic groups due to the in-group advantage [10]. However, results obtained by Lorenzo-Trueba et al. (2012) showed that Spanish subjects—the in-group listeners—could only identify two emotions more accurately: sadness at 78.10% and anger at 71.20%. Though Spanish listeners in our study could recognize *neutral* state with good accuracy (76.30%), Vietnamese listeners could identify the *neutral* state better at 84.00%. The fact that some groups from different Spanish regions express emotions more or less intensely [32] may explain the lower perception accuracies. For example, Spanish persons from Castilla are often thought to be more "serious" and "dramatic" whereas those from other regions may rarely use neutral tones as that is thought to be too serious [33]. Therefore, significant differences in results might be projected when subjects from different regions are asked to categorize the emotions.

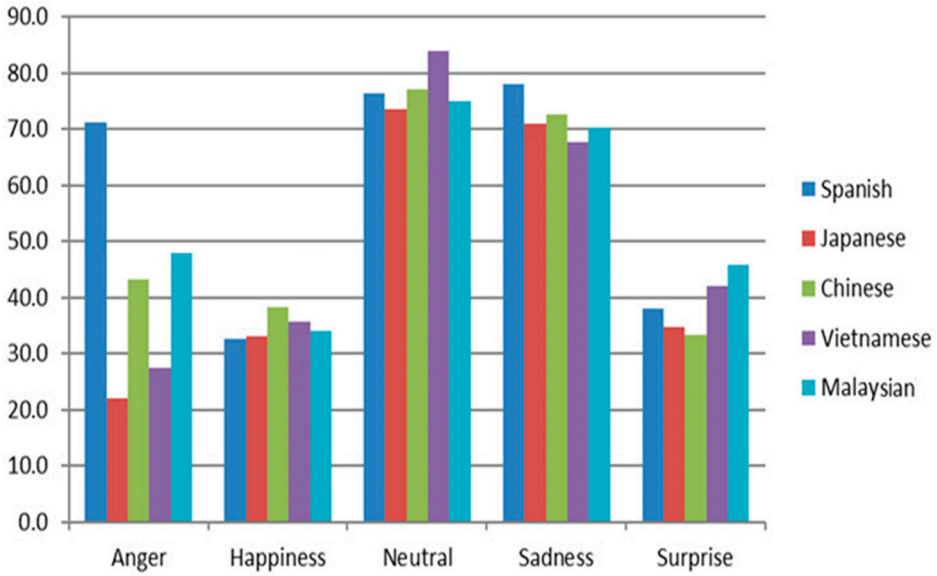

Figure 2. Bar chart representing percentage of Actual Emotion Perceived by Spanish, Japanese, Chinese, Vietnamese and Malaysian Subjects (based on the listening test using the SEV Corpus).

The results also showed some level of confusion among Spanish listeners when classifying *happiness* and *surprise* emotions, which were identified with much lower recognition rates at 32.70% and 38.10%, respectively. This could be due to the limitations of the synthetic speech, that produced a single joint emotion for happy surprise emotions where the Spanish listeners might have thought to be listening to happy surprised voices [30]. Attributes of both the emotions are high-pitched and lay closely in the valence-arousal emotional space, which may have caused the confusion [34]. Similarly, confusion while perceiving *anger*, *neutral* and *sadness* could be also influenced by the intensity of the utterances (from the SEV) which might have been inconsistent. This may be caused by the synthesized speech which is mostly distinguished by a lower accuracy, unnatural prosody and may lack the required expressiveness [35].

3.1.2. Out-Group Perception Results (Asians)

The Asian listeners perceived *neutral* and *sadness* at the highest recognition rates. They mainly perceived *anger* as *neutral*, *happiness* as *anger or surprise*, and *neutral* as *sadness or happiness*. Although the Asians showed almost similar tendencies in perceiving the different emotions and confusion (refer to Table 1), significant differences between the ethnic groups were noted:

- In general, *anger* is less acceptable in Asian cultures [36]; this is reflected by the perception rates of *anger* in the present experiment where Asians significantly perceived *anger* less than the Spanish. The lowest perception rate of *anger* was scored by the Japanese listeners, which is statistically significantly less than the Chinese ($p = 0.002$) and Malaysians (at $p < 0.001$). This finding is in line with findings of previous studies, where Japanese typically tend to avoid negative outcomes and thus engage in fewer negative emotions [37,38]. Evidence suggests that generally Asians suppress their negative emotions when perceiving out-group emotions, but not positive emotions such as *happiness* [39]. In this study, it was noted that all Asian groups perceived *happiness* higher than the Spanish listeners.
- Noticeably, Malaysians perceived Spanish *surprise* and *anger* more accurately than the other Asian ethnics. They perceived *surprise* statistically significantly higher than Japanese (at $p < 0.05$) and Chinese (at $p = 0.001$), whereas *anger* was perceived statistically significantly higher than the Japanese ($p < 0.001$) and Vietnamese ($p < 0.005$). Studies by Hei et al. [40] showed that Malaysians tend to be

more expressive in showing high-arousal emotions such as *surprise* and *anger* compared to other Asian ethnicities.

3.2. Analysis on Cultural Perception-Duration of Speech

One-way MANOVA analysis with the four durations as factors and the four emotions as dependent variables showed a statistically significant difference between the different groups of durations ($F(12, 87.6) = 2.37$, $p = 0.01$, Wilk's $\Lambda = 0.475$, partial $\eta 2 = 0.220$). Figure 3 illustrates the overall mean for emotion recognition rate by four Asian ethnic groups in four different durations. Asians found it difficult to perceive emotions in utterances with a duration longer than 0.04 s. A post-hoc analysis revealed that at duration 0.06 s, *happiness* and *sadness* were perceived statistically significantly lower than those perceived in other durations ($p < 0.05$). *Surprise* was also perceived to be statistically significantly lower at duration 0.06 s than at duration 0.02 s. The four emotions were perceived best at different durations: *happiness* at duration 0.03 s with a perception mean of 39.33 (a mark slightly higher than the perception mean at duration 0.04 s (39.15)), *sadness* at duration 0.04 s (75.83), *surprise* at duration 0.02 s (42.19), and *anger* at 0.02 s (37.61).

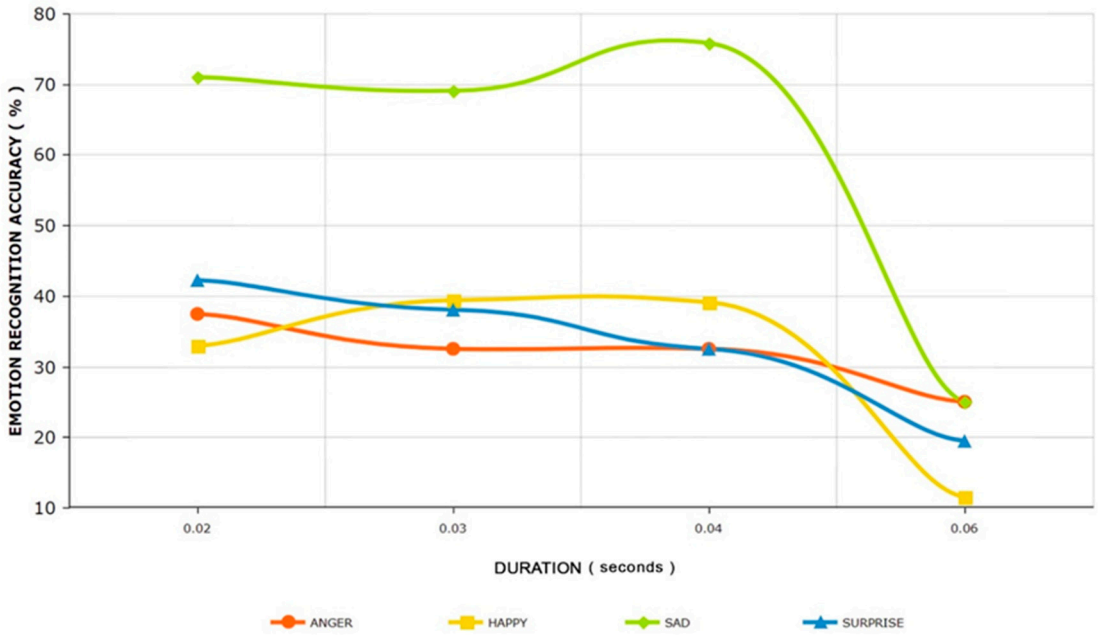

Figure 3. The overall mean for emotion perception by Asian ethnics based on speech duration.

3.3. Analysis on Cultural Perception-Intensity of Speech

A one-way MANOVA analysis was conducted to explore the significance of differences between the three intensities in perceiving the four emotions. The results showed statistically significant differences between the different intensities ($F(8, 68) = 37.45$, $p < 0.001$, Wilk's $\Lambda = 0.034$, partial $\eta 2 = 0.815$). Figure 4 shows that higher intensity could yield a better perception rate. A post-hoc analysis reported that all the emotions were perceived correctly at intensity 1 and intensity 1.25 decibels statistically significantly higher than at intensity 0.5 decibel (at $p \leq 0.001$). Though the perception means at intensity 1.25 decibels were higher for *sadness* and *surprise* and lower for *happiness* than perception means at intensity 1 decibel, the differences were not significant. Only *anger* was perceived correctly at intensity 1.25 decibels statistically significantly higher than at intensity 1 decibel ($p < 0.05$). In general, the results projected that the higher intensity gave a higher perception rate for all ethnic groups.

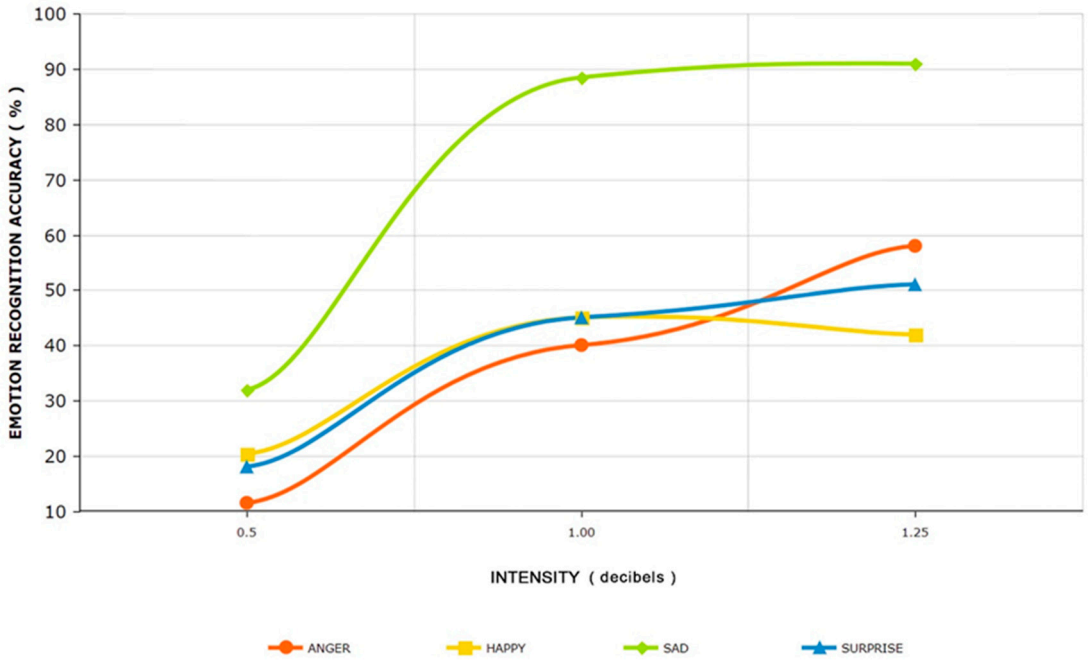

Figure 4. The overall mean for emotion perception by Asian ethnics based on speech intensity.

3.4. Analysis on Cultural Perception-Interaction between Ethnic, Duration and Intensity

A two-way MANOVA analysis was conducted on the different independent variables to investigate the effect of different combinations on perceiving emotions. The analysis revealed no significant interaction effect between intensity and duration or ethnic and duration variables, but did so between ethnic and intensity variables. A statistically significant ethnic x intensity multivariate main effect was also noted on perceiving the different emotions (F (24, 88.43) = 2.19, p = 0.004, Wilks' Λ = 0.196). The univariate analysis revealed that the significant effect was in perceiving *surprise* (at p = 0.001) and *anger* ($p < 0.001$). The means in Figure 5 indicate that although the Malaysians perceived *surprise* correctly higher than the Vietnamese and statistically significantly higher than the Japanese ($p < 0.05$) and Chinese ($p < 0.001$) at intensity 0.5 and intensity 1 decibels, they perceived *surprise* less at intensity 1.25 decibels. Other ethnic groups showed a direct proportion between the intensity and the mean of perceiving emotion.

The results illustrated in Figure 5 show that intensity of speech could lead to inaccurate perception by out-group listeners. Experiments by Lee et al. [41] found that, in Mandarin Chinese, syllables were lower in pitch and intensity. The Japanese language and similar languages such as Chinese and Vietnamese language, have vowel lengths that are phonemic such that shortening or lengthening spoken vowels changes word meanings [42]. Hence, the different features of utterances in Spanish may have influenced the perception rate among the different ethnic groups. The different melodic and acoustic patterns in languages spoken by Asians could be one of the main reasons of variation in results for the duration and intensity factors. These factors should be taken as a main consideration during the implementation of conversational artificial agents in any domain. Figure 5 shows the significant effect of intensity in perceiving anger by the Chinese ethnic group. Although the Chinese recorded the lowest perception mean of anger at intensity 0.5 decibel, they perceived the same emotion better than the other races at intensity 0.5 and intensity 1 decibels.

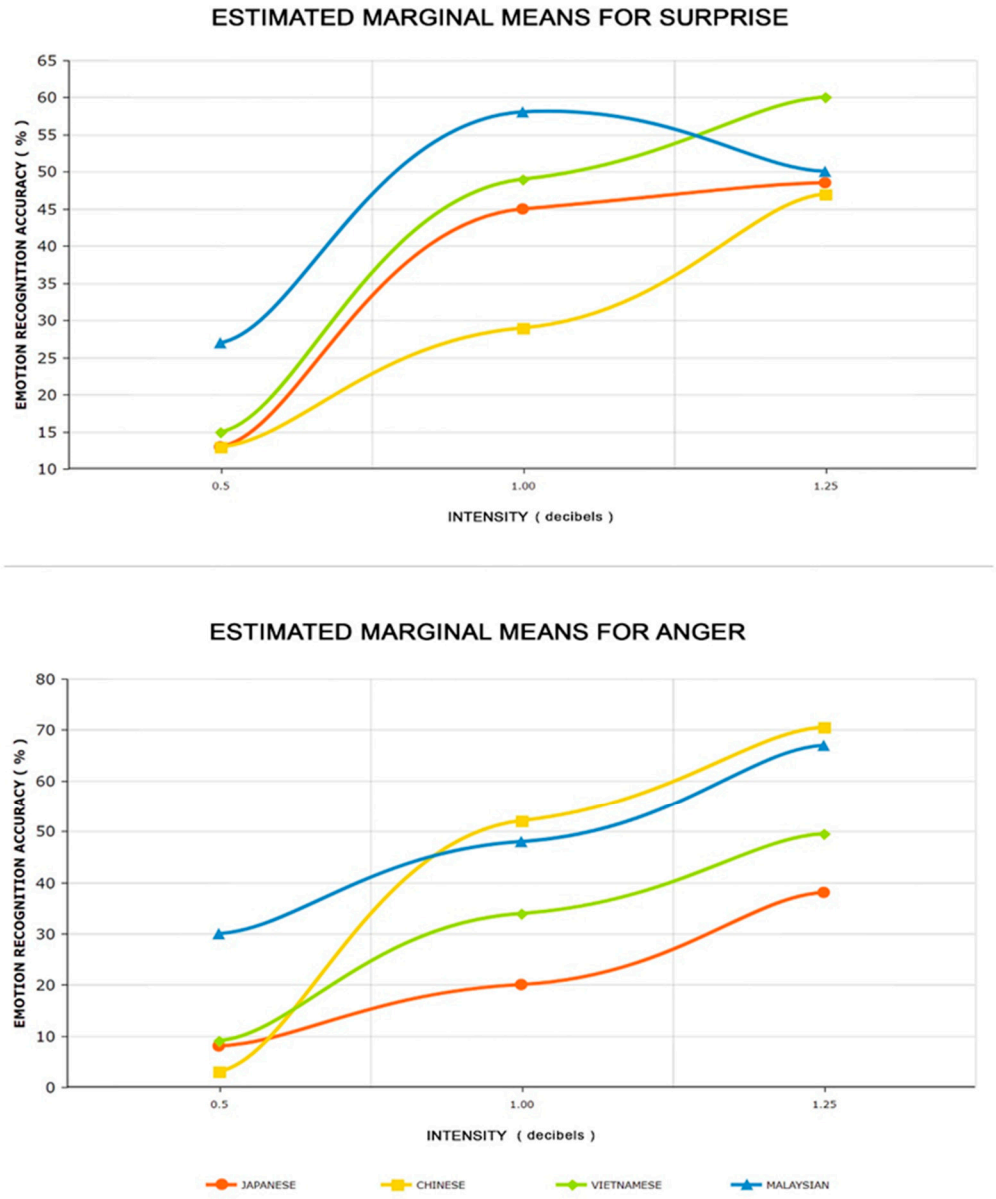

Figure 5. Graph showing the two-way MANOVA analysis between Intensity and Ethnicity for Anger and Surprise emotions.

4. Conclusions

Speech contains rich information beyond what is said, such as the speaker's emotion. Accurately recognizing these emotions is important when developing effective human–machine applications. Translation of speech elicited from one culture may not be perceived correctly by other cultures and may lead to misperception.

Prosody (intensity, vocal pitch, rhythm, rate of utterance) in speech plays a major role in expressing emotions and can be intentionally modified to communicate different feelings [43]. In this paper, we investigated the expressive synthetic voices cross which was perceived by different ethnic groups using two different features (duration and intensity). Results show similarities as well differences in the perception rate among ethnics. Thus, it is apparent that speech features, such as duration and intensity, play a vital role in order to project the emotions as intended, especially in a cross-cultural setting. Careful consideration of these features is essential in developing a model for agents in real-time application which supports the social nature of culture-specific interactions in a cross-cultural setting.

In our future work, we would like to extend the experiment using different corpora and ethnic groups in order to obtain more detailed findings on human perception and other features that should be considered when modelling artificial agents.

Acknowledgments: Authors are thankful to School of Computer Sciences, Universiti Sains Malaysia and the Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Spain for the support s supported by short-term grant: 304/PKOMP/6312153 awarded by. The authors would like to acknowledge the support of Masato Akagi, Reda Elbarogy, members of Unoki & Akagi Lab, Japan Advanced Institute of Science and Technology (JAIST).

Author Contributions: Ganapreeta Renunathan Naidu performed the experiments and collected the data, analyzed the data and wrote the paper. Amal Abdulrahman Azazi performed statistical analysis and assisted in writing the paper. Syaheerah Lebai Lutfi obtained the Spanish expressive voice (SEV) corpus, directed this research work and reviewed the paper contents to improve the manuscript, to make it a worthy contribution to the scientific community. Juan Manuel Montero Martinez designed and labelled the SEV corpus and Jaime Lorenzo-Trueba validated it.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Remland, M.S. *Nonverbal Communication in Everyday Life*; SAGE Publications: Thousand Oaks, CA, USA, 2017.
2. Warren, P. Prosody and language processing. In *Language Processing*; Psychology Press Ltd.: London, UK, 1999; pp. 155–188.
3. Banse, R.; Scherer, L. Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* **1996**, *70*, 614–636. [CrossRef] [PubMed]
4. Juslin, P.N.; Laukka, P. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychol. Bull.* **2003**, *129*, 770–814. [CrossRef] [PubMed]
5. Scherer, K.R. Vocal communication of emotion: A review of research paradigms. *Speech Commun.* **2003**, *40*, 227–256. [CrossRef]
6. Graham, C.R.; Hamblin, A.; Feldstein, S. Recognition of emotion in English voices by speakers of Japanese, Spanish, and English. *IRAL* **2001**, *39*, 19–37. [CrossRef]
7. Thompson, W.F.; Balkwill, L.-L. Decoding speech prosody in five languages. *Semiotica* **2006**, *158*, 407–424. [CrossRef]
8. Bryant, G.A.; Barrett, H.C. Vocal emotion recognition across disparate cultures. *J. Cognit. Cult.* **2008**, *8*, 135–148. [CrossRef]
9. Pell, M.D.; Paulmann, S.; Dara, C.; Alasseri, A.; Kotz, S.A. Factors in the recognition of vocally expressed emotions: A comparison of four languages. *J. Phon.* **2009**, *37*, 417–435. [CrossRef]
10. Elfenbein, H.A.; Ambady, N. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychol. Bull.* **2002**, *128*, 203–235. [CrossRef] [PubMed]
11. Barra Chicote, R.; Montero Martínez, J.M.; Macías Guarasa, J.; Lutfi, S.L.; Lucas Cuesta, J.M.; Fernández Martínez, F.; D'haro Enríquez, L.F.; San Segundo Hernández, R.; Ferreiros López, J.; Córdoba Herralde, R.D.; et al. Spanish expressive voices: Corpus for emotion research in Spanish. In Proceedings of the 6th Conference of Language Resources and Evaluation (Workshop on Corpora for Research on Emotion and Affect), Marrakech, Morocco, 26 May 2008.
12. Katsumi, Y.; Kim, S.; Sung, K.; Dolcos, F.; Dolcos, S. When nonverbal greetings "Make it or break it": The role of ethnicity and gender in the effect of handshake on social appraisals. *J. Nonverbal Behav.* **2017**, *41*, 345–365. [CrossRef]
13. Scherer, K.R. Vocal affect expression: A review and a model for future research. *Psychol. Bull.* **1986**, *99*, 143–165. [CrossRef] [PubMed]
14. Wilson, D.; Wharton, T. Relevance and prosody. *J. Pragmat.* **2006**, *38*, 1559–1579. [CrossRef]
15. Root, A.R. The pitch factors in speech—A survey. *Q. J. Speech* **1930**, *16*, 320–343. [CrossRef]
16. Pluggé, D.E. "Voice qualities" in oral interpretation. *Q. J. Speech* **1942**, *28*, 442–444. [CrossRef]
17. Lewis, M.; Takai-Kawakami, K.; Kawakami, K.; Sullivan, M.W. Cultural differences in emotional responses to success and failure. *Int. J. Behav. Dev.* **2010**, *34*, 53–61. [CrossRef] [PubMed]

18. Beier, E.G.; Zautra, A.J. Identification of vocal communication of emotions across cultures. *J. Consult. Clin. Psychol.* **1972**, *39*, 166. [CrossRef] [PubMed]
19. Jacewicz, E.; Fox, R.A.; O'Neill, K.; Salmons, J. Articulation rate across dialect, age, and gender. *Lang. Var. Chang.* **2009**, *21*, 233–256. [CrossRef] [PubMed]
20. Adank, P.; Janse, E. Perceptual learning of time-compressed and natural fast speech. *J. Acoust. Soc. Am.* **2009**, *126*, 2649–2659. [CrossRef] [PubMed]
21. Heald, S.; Klos, S.; Nusbaum, H. Understanding Speech in the Context of Variability. In *Neurobiology of Language*; Elsevier Science: Amsterdam, The Netherlands, 2015; pp. 195–208.
22. McCulloch, G. The very idea of the phenomenological. *Proc. Aristot. Soc.* **1993**, *93*, 39–57. [CrossRef]
23. Bassiouney, R. (Ed.) *Identity and Dialect Performance: A Study of Communities and Dialects*; Routledge: Abingdon, UK, 2017.
24. Iacobelli, F.; Cassell, J. Ethnic identity and engagement in embodied conversational agents. In *Intelligent Virtual Agents*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 57–63.
25. Cassell, J. Social practice: Becoming enculturated in human-computer interaction. In *Universal Access in Human-Computer Interaction. Applications and Services*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 303–313.
26. Sharma, L.; Srivastava, V. Performance enhancement of information retrieval via Artificial Intelligence. *IJSRSET* **2017**, *3*, 187–192.
27. Reeves, B.; Nass, C. *How People Treat Computers, Television, and New Media Like Real People and Places*; CSLI Publications Stanford University: Stanford, CA, USA, 1996; pp. 3–18.
28. Nass, C.I.; Brave, S. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*; MIT Press: Cambridge, MA, USA, 2005.
29. Lorenzo-Trueba, J.; Watts, O.; Barra-Chicote, R.; Yamagishi, J.; King, S. Simple4all proposals for the albayzin evaluations in speech synthesis. In Proceedings of the IberSPEECH 2012, Madrid, Spain, 21–23 November 2012.
30. Barra-Chicote, R.; Yamagishi, J.; King, S.; Montero, J.M.; Macias-Guarasa, J. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Commun.* **2010**, *52*, 394–404. [CrossRef]
31. *SPSS [Computer Software]*; Version 11.5; SPSS Inc.: Chicago, IL, USA, 2013.
32. Elfenbein, H.A.; O'Reilly, C.A. Fitting in: The effects of relational demography and person-culture fit on group process and performance. *Group Organ. Manag.* **2007**, *32*, 109–142. [CrossRef]
33. Petkova, D. Cultural Diversity in People's Attitudes and Perceptions. April 2006. Fondazione Eni Enrico Mattei Working PAPER No. 56.2006. Available online: http://dx.doi.org/10.2139/ssrn.897423 (accessed on 16 January 2017).
34. Jones, C.M.; Jonsson, I.M. Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, Canberra, Australia, 21–25 November 2005; pp. 1–10.
35. Iriondo, I.; Guaus, R.; Rodríguez, A.; Lázaro, P.; Montoya, N.; Blanco, J.M.; Bernadas, D.; Oliver, J.M.; Tena, D.; Longhi, L. Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, UK, 5–7 September 2000.
36. Paulmann, S.; Uskul, A.K. Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognit. Emot.* **2014**, *28*, 230–244. [CrossRef] [PubMed]
37. Zahn-Waxler, C.; Friedman, R.J.; Cole, P.M.; Mizuta, I.; Hiruma, N. Japanese and United States preschool children's responses to conflict and distress. *Child Dev.* **1996**, *67*, 2462–2477. [CrossRef] [PubMed]
38. Mesquita, B.; Karasawa, M. Different emotional lives. *Cognit. Emot.* **2002**, *16*, 127–141. [CrossRef]
39. Hurley, C.M.; Teo, W.J.; Kwok, J.; Seet, T.; Peralta, E.; Chia, S.Y. Diversity from within: The Impact of Cultural Variables on Emotion Expressivity in Singapore. *IJPS* **2016**, *8*, 50. [CrossRef]
40. Hei, K.C.; Ling, W.N.; David, M.K. Communicating Disagreements among Malaysians: Verbal or Non-verbal? *Lang. India* **2011**, *11*, 442–462.
41. Lee, Y.C.; Wang, T.; Liberman, M. Production and Perception of Tone 3 Focus in Mandarin Chinese. *Front. Psychol.* **2016**. [CrossRef] [PubMed]

42. Hirata, Y. Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *J. Acoust. Soc. Am.* **2004**, *116*, 2384–2394. [CrossRef] [PubMed]
43. Lutfi, S.L.; Fernández-Martínez, F.; Lorenzo-Trueba, J.; Barra-Chicote, R.; Montero, J.M. I feel you: The design and evaluation of a domotic affect-sensitive spoken conversational agent. *Sensors* **2013**, *13*, 10519–10538. [CrossRef] [PubMed]