Analysis of Mandarin *vs* English Language for Emotional Voice Conversion

S. Uthiraa(✉) and Hemant A. Patil

Speech Research Lab, DA -IICT, Gandhinagar, Gujarat, India
{uthiraa_s,hemant_patil}@daiict.ac.in

Abstract. Emotional Voice Conversion (EVC) is a method to convert the emotional state of an utterance to another without changing the linguistic information and speaker's identity. Its application is enormous in human-machine interaction, development of emotional Text-To-Speech (TTS), etc. This study focuses on analyzing the characteristics of Mandarin and English language for EVC between these languages. Prosodic features, such as energy, fundamental or pitch frequency (F_0), duration, pauses/silences, and loudness are compared using several techniques, such as narrowband spectrograms, Root Mean Square Energy (RMSE), and Zero-Crossing Rate (ZCR). Teager Energy Operator (TEO) based features are studied to analyze the energy profile of emotions. The Emotional Speech Dataset (ESD) is used in this work. Experiments were performed on 5 emotions, namely, anger, happiness, neutral, sad, and surprise. Results showed that tonal language (i.e., Mandarin) has steep and multiple fluctuations in F_0 contour as it is pitch-dependent, as compared to the stress-time language (English), which had less F_0 fluctuations, and is stable for the most duration of the sentence. Loudness and silences are also different in the two languages. These findings may serve as important cues for EVC task.

Keywords: Emotional Voice Conversion · Emotional Speech Database (ESD) · Narrowband Spectrogram · Fundamental Frequency · Teager Energy Operator

1 Introduction

Communication, "the mode for transferring, sharing, and receiving information", which is performed by either verbal, non-verbal or visual means. Language, "a structured system of communication", conveyed through speech (spoken), writing or signs. In this paper, we focus on the spoken aspect of language. In this era, where population and technology is increasing rapidly, communication among and between them is essential. Language plays its role well for human interaction as well as for human-machine interaction. Moreover, language is the engine of cultivation and human speech is its most powerful form.

Voice Transformation (VT) aims at changing one or more aspects of a speech signal while preserving its linguistic information. Voice Conversion (VC) aims at

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 295–306, 2023.
https://doi.org/10.1007/978-3-031-48312-7_24

changing ***source*** speaker's voice in such a way that, it sounds as if the ***target*** speaker has spoken that sentence [7]. In this context, Emotional Voice Conversion (EVC) aims to convert the emotional state of the utterance, while preserving the linguistic and speaker information [14]. This paper focuses on analysis of emotions in Mandarin ***vs.*** English in the context of EVC as it has significant application in human-machine interaction [9], and aids at developing emotional Text-To-Speech (TTS).

The earlier work on EVC dates back to around 2003 [5], where neutral speech was converted to other emotions, such as joy, anger, happiness, etc. For emotion recognition, one of the prominent features is prosodic feature extraction, which includes tone, rhythm, intonation, energy, duration, fundamental frequency (F_0), and loudness parameters [10]. For this paper, we use prosodic features, such as energy, loudness, F_0 to compare the emotions produced in Mandarin and English languages. This feature is selected as Mandarin is known to be a ***tonal*** language and English is a stress-timed language and thus, prosodic features will aid in its analysis [12].

In this paper, we analyze five emotions, namely, anger, happy, neutral, sad, and surprise in English and Mandarin language using narrowband spectrograms, F_0, Root Mean Square Energy (RMSE) and Zero-Crossing Rate (ZCR) to investigate prosodic parameters that are essential and more significant for emotional voice conversion between languages. Observations indicate that RMS and ZCR values can be used for EVC between languages.

The rest of the paper is organized as follows: In Sect. 2, we discuss the proposed work. Section 3 gives the details of the experimental setup. Section 4 presents the analysis of the results. Section 5 concludes the paper along with potential future research directions.

2 Proposed Work

Several languages in Southeast Asia and Africa are tonal languages, where pitch or F_0 differences are used to differentiate meanings of words or to convey grammatical distinctions. In contrast, English is a stress-timed language, i.e., in this language, the tone is used to convey an attitude or change a statement to a question, however, it does not affect the meaning of individual words [1].

In the baseline paper [13], EVC was performed in the same language, i.e., English neutral was converted to English sad or happy. The analysis presented in this paper is useful for conversion between languages and between emotions. In this paper, we analyze the loudness parameter using RMSE, voiced and unvoiced components using ZCR, and F_0 and its harmonics using narrowband spectrograms.

2.1 Spectrographic Analysis

Spectrograms are a visual representation of acoustic signals with time (X-axis), frequency (Y-axis), and amplitude measures in parameter representation. Pauses

and harmonic components are also seen. In this paper, we study the narrowband spectrograms (as they give good frequency resolution, i.e., show pitch source harmonics as horizontal striations, useful for tonal language analysis), and F_0 of English and Mandarin sentences spoken in 5 emotions, namely, anger, happy, neutral, sad, and surprise. The energy distribution, pitch source harmonics, and silences are compared. Figures 1 and 2 shows the F_0 changes, plot, and spectrograms of female speakers uttering the same sentence in English and Mandarin, respectively.

2.2 Root Mean Square (RMS) Energy

RMS for speech signal is a crucial acoustic cue for target speech perception [11]. It is the squared signal value (amplitude), averaged over time, and its square root is calculated. In particular,

$$RMS_t = \sqrt{1/K \sum_{n=t.K}^{(t+1)(K-1)} |s(n)^2|}, \tag{1}$$

where $s(n)^2$ is the energy of n^{th} sample, then we sum the energies of all the samples at time $\boldsymbol{t}$. To get the mean, it is then divided by frame size, K.

This feature has significant applications in audio segmentation and music genre classification. In this paper, we plot the RMS values of audio to find the loudness measure. Amplitude envelope (AE) can also be used to measure loudness, however, RMS is preferred as it is less sensitive to outliers than the AE. In addition, it gives us perceived loudness, i.e., the way our ear perceives loudness. In Fig. 3, each plot depicts the RMS values of the same sentences spoken in English (yellow colored) and Mandarin (Red colored) by 2 female (1 for English and 1 for Mandarin) speakers in 5 emotions, namely, anger, happy, neutral, sad, and surprise, respectively.

2.3 Zero-Crossing Rate (ZCR)

ZCR is "the rate at which a signal changes from positive to zero to negative or from negative to zero to positive". Historically, it is known to have a correlation with formants, thus, helpful for speech perception [6]. Its expressed as-

$$ZCR_t = (1/2). \sum_{n=t.K}^{(t+1)(K-1)} |sgn(s(n)) - sgn(s(n+1)), \tag{2}$$

where s(n) and s(n+1) represent the amplitude at sample n and its consecutive amplitude sample, respectively.

It is an useful measure to recognize percussive (random ZCR) ***vs.*** pitched sounds (stable ZCR) [4]. For this work, we use ZCR for monotonic pitch estimation and for analyzing the voiced and unvoiced segments of audio signal [3]. Figure 4 shows the ZCR plot for 2 females (1 for English and 1 for Mandarin) speaking the same sentence in both languages with 5 emotions, namely, anger, happy, neutral, sad, and surprise, respectively.

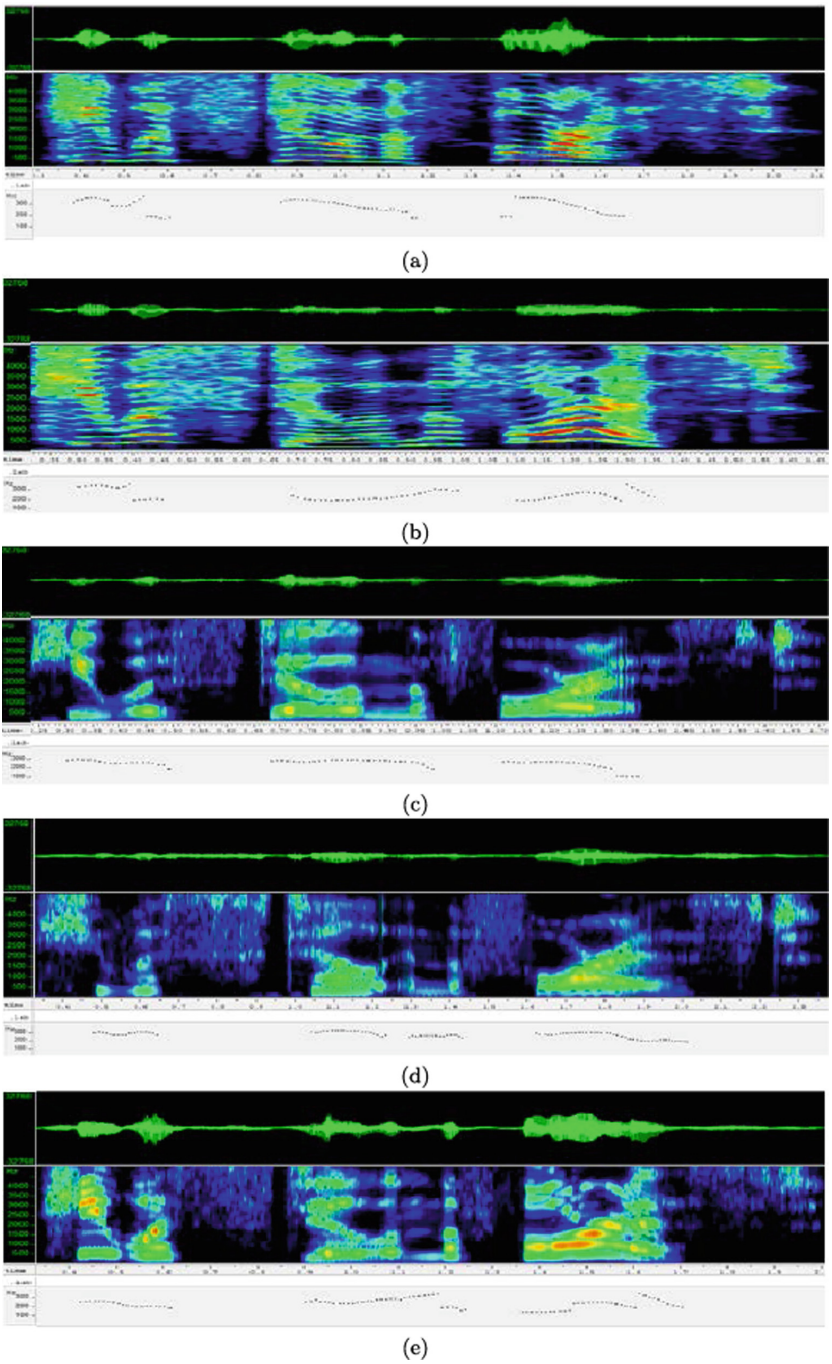

Fig. 1. Time-domain signal, narrowband spectrograms, F_0 contour of English sentences by female speakers in 5 emotions: (a) anger, (b) happy, (c) neutral, (d) sad, and (e) surprise.

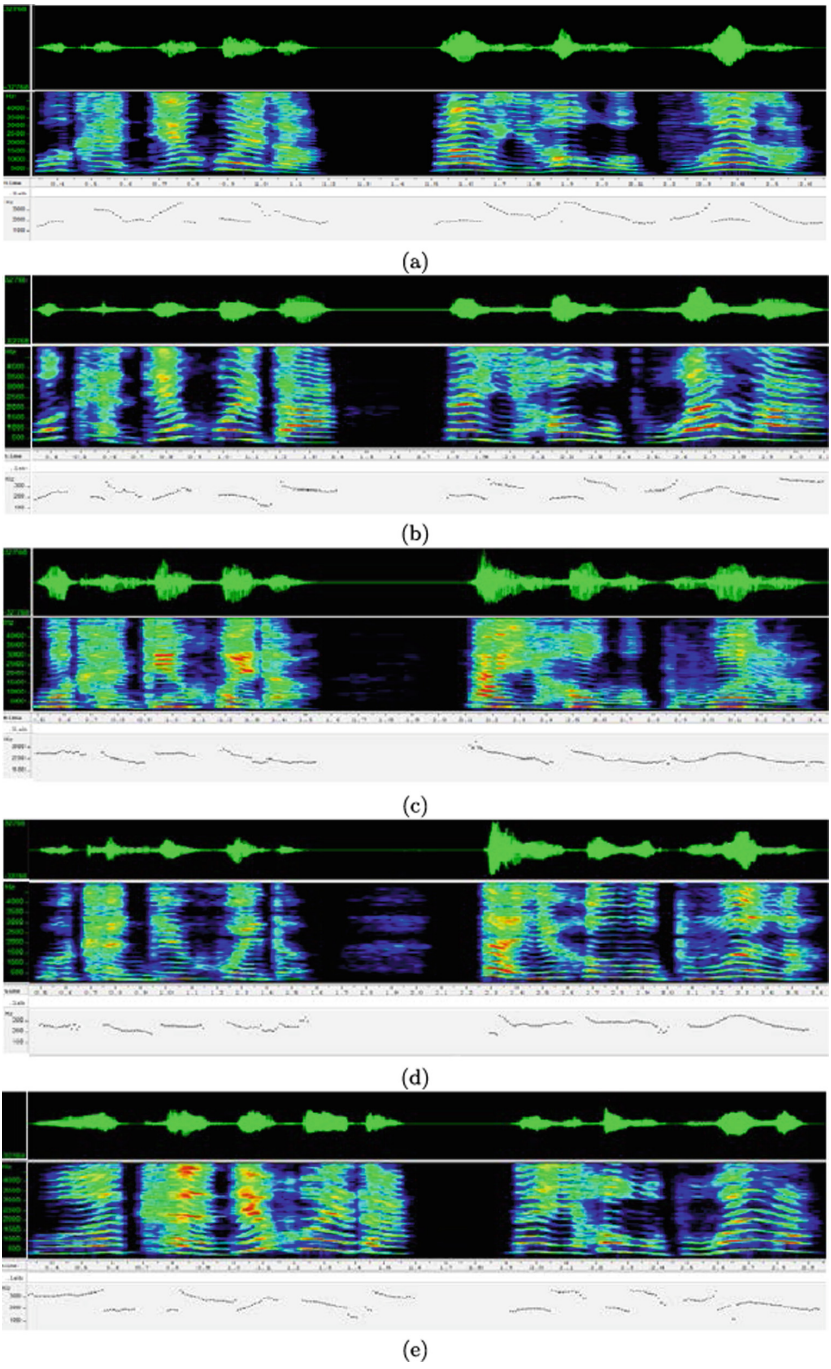

Fig. 2. Time-domain signal, narrowband spectrograms, F_0 contour of Mandarin sentences by female speakers in 5 emotions: (a) anger, (b) happy, (c) neutral, (d) sad, and (e) surprise.

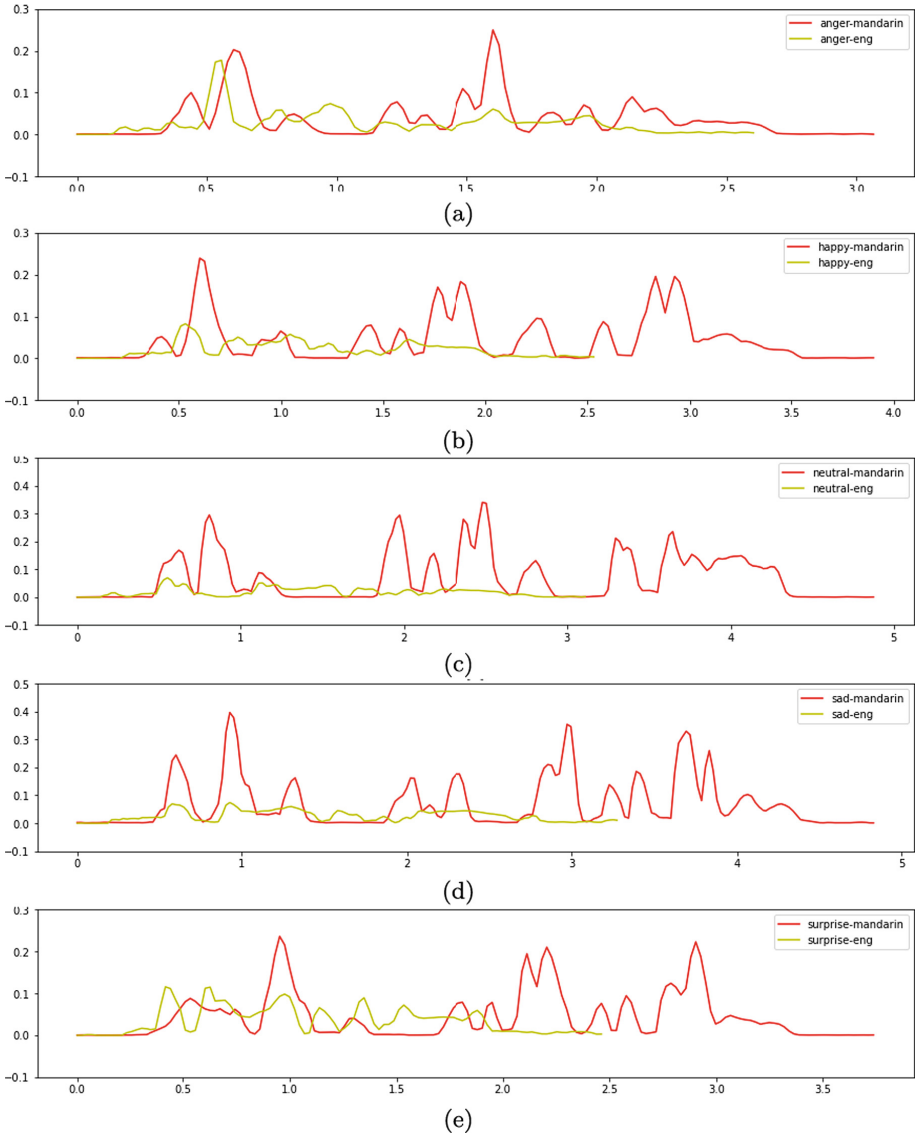

Fig. 3. RMS for Mandarin ***vs.*** English for a sentences in (a) anger, (b) happy, (c) neutral, (d) sad, and (e) surprise by female speakers.

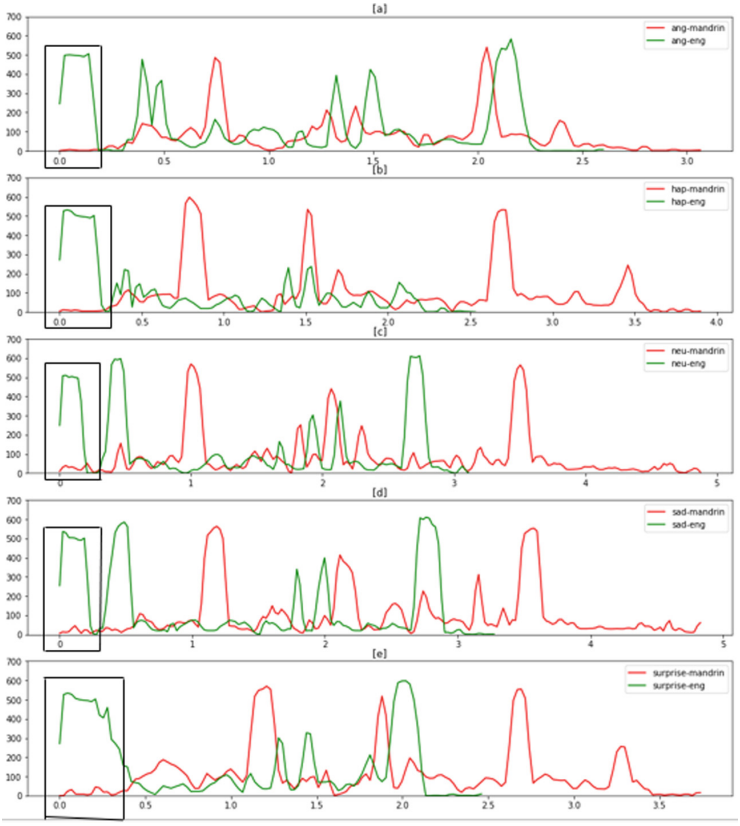

Fig. 4. ZCR for Mandarin ***vs.*** English for a sentences in [a] anger, [b] happy, [c] neutral, [d] sad, and [e] surprise by female speakers. The box at the beginning of the plot indicates the whisper sound —h— in "he" uttered.

2.4 Teager Energy Operator (TEO)

Speech is produced by non-linear, vortex airflow interaction in the vocal tract. A stressful situation affects the muscle tension of the speaker which results in an alteration of the airflow during the production of the sound [2]. This is captured ***via*** TEO, in particular, $\Psi\{x(n)\} = x^2(n) - x(n+1)x(n-1)$, where $\Psi\{\}$ is the Teager Energy Operator (TEO), and $x(n)$ is the discrete-time signal. TEO features are extensively used in distinguishing genuine ***vs.*** replay speech in spoofing. In this paper, we use TEO to analyze the glottal closure impact, i.e., bumps within the glottal cycle are studied [8]. Figures 5 and 6 have the TEO profile of a female speaker uttering the same sentence with 5 emotions in English and Mandarin, respectively, with the X-axis representing frames and the Y-axis, amplitude. Figures 5 and 6 show that the TEO gives a running estimate of the signal's energy w.r.t. time. Further, the TEO profile seems to vary across emotions for a particular language (here, either Mandarin or English).

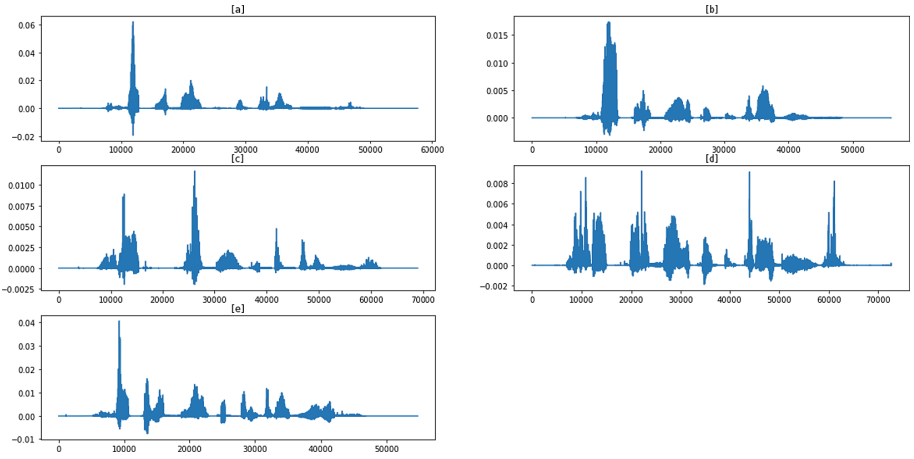

Fig. 5. TEO profile of a female speaker uttering an English sentence in [a] anger, [b] happy, [c] neutral, [d] sad, and [e] surprise.

3 Experimental Results

3.1 Dataset Used

In this paper, we have used a recently developed ESD dataset [13]. It consists of 350 parallel utterances spoken by 10 native English (5 female and 5 male), and 10 native Mandarin speakers (5 female and 5 male) speakers. The emotions captured in it are - anger, happy, neutral, sad, and surprise, whose audio is sampled at 16 kHz. This dataset is chosen as it is a relatively large-scale, multi-speaker and publicly available dataset with good recording conditions [14], thus, making the analysis relatively accurate.

3.2 Experimental Results

All the results mentioned are generalized results which were taken and compared with atleast 5 sentences for each emotion, but for the paper readability, results using only 1 sentence (from female speakers) are given. The analysis for male speakers was similar to that of female speakers, but the distinction between emotions was clearer for females than males. The detailed analysis of spectrograms (shown in Figs. 1 and 2) is presented in Fig. 7. We infer that high energy contents are seen in all 5 emotions of Mandarin speech and thus, indicating that Mandarin speech is usually louder in comparison to English speech. A significant difference seen in spectrograms is that all English sentences with 5 emotions had energy components present only at the higher frequency at the end of a sentence, which wasn't seen in any spectrograms for Mandarin. The width between the two consecutive horizontal striations in the narrowband spectrogram gives pitch

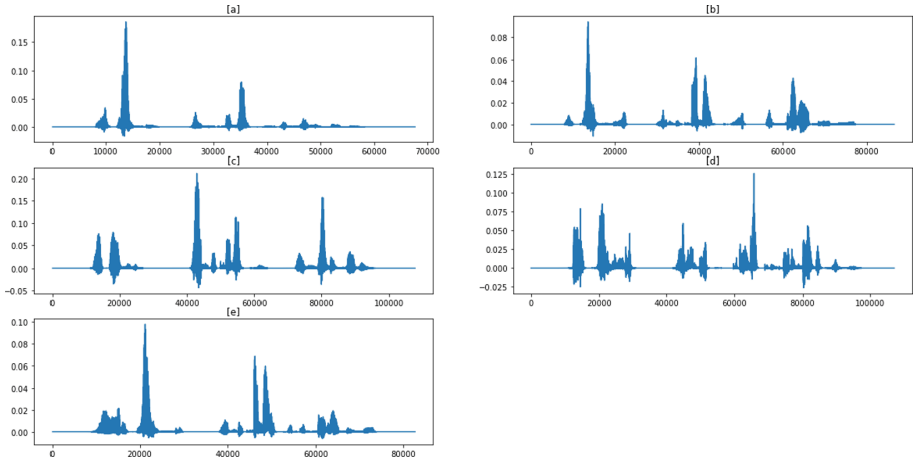

Fig. 6. TEO profile of a female speaker uttering a Mandarin sentence in [a] anger, [b] happy, [c] neutral, [d] sad, and [e] surprise.

(the way the auditory system perceives frequency) information, which is higher in Mandarin than in English. The silences were seen more in Mandarin than in English.

The study of F_0 contour is represented in the form of a boxplot (which gives the spread or variance of F_0) in Fig. 8. It is noted that neutral emotion has the least spread in both languages and the highest spread is seen in emotions; surprise and anger in English and Mandarin speech, respectively. Almost no outliers are seen for Mandarin speech, i.e., there is not much difference between the F_0 values as compared to English. Another distinction seen is that the median values for all emotions in Mandarin are higher than that in English. These conclude that the F_0 contours are at higher frequencies, and with wide fluctuations for Mandarin speech.

In the RMS plots (Fig. 3), it is observed that all the emotional sentences spoken in Mandarin has significant fluctuations in peaks compared to the English statements. Anger and surprise emotions have similar peaks in both the languages. Neutral and sad sentences in English have almost no variations in peaks. Happy in Mandarin has broader peaks. These results state that Mandarin sentences are perceived louder (as have more energy content, as seen from spectrograms) than the corresponding English sentences.

Characteristics	Anger	Happy	Neutral	Sad	Surprise
Energy Content	High energy content is not uniformly distributed in English as compared to Mandarin.	High energy content is more prominent in lower frequencies in Mandarin but also seen at higher frequencies in the beginning of the sentence in English.	Very less high energy content in English but in Mandarin, high energy content is present.	Very less high energy content in English. In Mandarin, high-energy content seen towards the end of the sentence.	Almost similar high energy content seen in both English and Mandarin.
Width	Width between harmonics is higher in Mandarin than in English.	Width between harmonics is more in Mandarin than in English.	Width between harmonics is more in Mandarin than in English.	Width between harmonics is more in Mandarin than in English.	Width between harmonics is more in Mandarin than in English.
Pauses	Clear and distinct in Mandarin than in English.	More are distinct pauses seen in Mandarin than in English.	Clear and low duration pauses in English. In Mandarin, clear and higher duration pauses seen.	Similar pauses seen in both English and Mandarin except for 1 long pause in Mandarin.	Distinct pauses seen in Mandarin.

Fig. 7. Analysis of narrowband spectrograms for English ***vs.*** Mandarin emotions.

The ZCR plots shown in Fig. 4, give the idea on percussive ***vs.*** pitched sounds. We can consider two extreme cases of spectral energy density, i.e., the low frequency and high frequency regions. It is observed that ZCR peaks are less in lower frequency regions and high in higher frequency regions of spectrograms. ZCR peaks of Mandarin are less than that of English as tonal sounds are pitch-dependent and have voiced speech as compared to English, which has unvoiced and whisper elements (beginning of the sentence, as shown in Fig. 4 for the sentence analyzed, and thus, proving that ZCR peaks are high for unvoiced sounds in comparison to their voiced counterpart).

The TEO plots in Figs. 5 and 6 show that Mandarin sentences have higher energy profiles (peaks reach higher amplitudes) than English sentences. This is because a higher pitch leads to higher loudness and thus, higher amplitude.

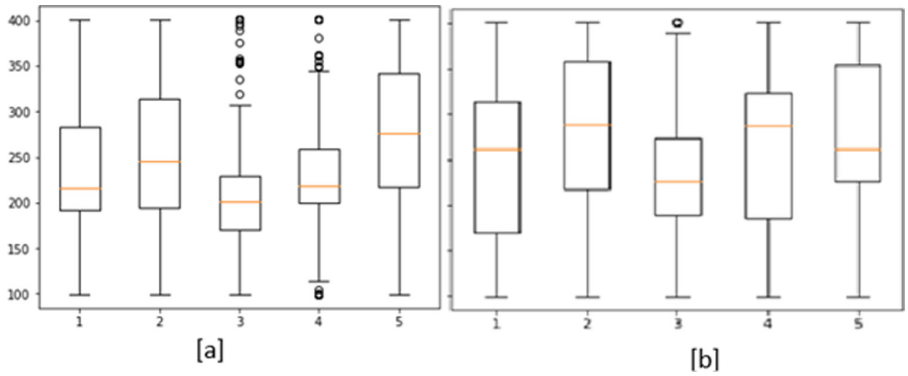

Fig. 8. Boxplot of F_0 contour of female speaker uttering an [a] English and [b] Mandarin sentence in [1] anger, [2] happy, [3] neutral, [4] sad, and [5] surprise.

4 Summary and Conclusion

In this study, we analyze a tonal language (Mandarin), and a stress-timed language (English) using prosodic features, such as energy, F_0, loudness, and TEO-based features. Our analysis indicate, Mandarin language has higher F_0 fluctuations due to variations in pitch, are louder, and have higher energy profiles than English language. Therefore, for EVC, RMS, and ZCR features can be used to maintain the speaker's identity. It would be interesting to analyze how RMS and ZCR features would work if, replaced with F_0 in the baseline paper [13] for EVC. The study presented in this paper may help in analyzing the confusion matrices that are obtained from the SER task. Future work includes using these results in classifiers for performing EVC in the same and in multi-languages and developing more datasets w.r.t. EVC.

Acknowledgements. The authors are thankful to the Ministry of Electronics and Information Technology (MeitY), New Delhi, Government of India, for sponsoring the project, "National Language Translation Mission (NLTM): BHASHINI with the objective of Building Assistive Speech Technologies for the Challenged (Grant ID: 11(1)2022-HCC (TDIL)).

References

1. An introduction to Tonal languages. https://ceas.sas.upenn.edu/sites/default/files/. Accessed 9 Sep 2022
2. Alex, S.B., Mary, L., Babu, B.P.: Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features. Circuits Syst. Signal Process. **39**(11), 5681–5709 (2020)
3. Bachu, R., Kopparthi, S., Adapa, B., Barkana, B.: Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In: American Society for Engineering Education (ASEE) Zone Conference Proceedings, pp. 1–7. American Society for Engineering Education (2008)

4. Brata, I., Darmawan, I.: Comparative study of pitch detection algorithm to detect traditional balinese music tones with various raw materials. J. Phys.: Conf. Ser. **1722**, 012071 (2021)
5. Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H., Shikano, K.: GMM-based voice conversion applied to emotional speech synthesis. In: 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland (2003)
6. Licklider, J.C.R., Pollack, I.: Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. J. Acoust. Soc. Am. **20**(1), 42–51 (1948)
7. Mohammadi, S.H., Kain, A.: An overview of voice conversion systems. Speech Commun. **88**, 65–82 (2017)
8. Patil, H.A., Parhi, K.K.: Development of TEO phase for speaker recognition. In: 2010 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IISc Bangalore, India (2010)
9. Pittermann, J., Pittermann, A., Minker, W.: Handling Emotions in Human-computer Dialogues. Springer (2010). https://doi.org/10.1007/978-90-481-3129-7
10. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. **21**(1), 93–120 (2018). https://doi.org/10.1007/s10772-018-9491-z
11. Wang, L., Wu, E.X., Chen, F.: Contribution of RMS-level-based speech segments to target speech decoding under noisy conditions. In: INTERSPEECH, pp. 121–124, Shangai China (2020)
12. Zhou, K., Sisman, B., Li, H.: Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. arXiv preprint arXiv:2002.00198 (2020)
13. Zhou, K., Sisman, B., Liu, R., Li, H.: Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, pp. 920–924 (2021)
14. Zhou, K., Sisman, B., Liu, R., Li, H.: Emotional voice conversion: theory, databases and ESD. Speech Commun. **137**, 1–18 (2022)