EVALUATION OF MIMICKED SPEECH USING PROSODIC FEATURES

Leena Mary, Anish Babu K. K, Aju Joseph, Gibin M. George
Rajiv Gandhi Institute of Technology, Kottayam, Kerala, India-686 501
leena.mary@rit.ac.in, anishkochi@yahoo.co.uk, aju.official@gmail.com, gibinmgeorge89@gmail.com

ABSTRACT

In this paper, we describe a technique for evaluating the quality of mimicked speech. In other words, mimicry artists are evaluated based on their competences to mimic a particular person. This evaluation is done based on prosodic characteristics for the text dependent cases. Prosodic characteristics are represented using features derived from pitch contour, duration and energy. In this work, prosodic features are extracted from speech after automatically segmenting into intonational phrases. Pitch contour corresponding to each phrase is approximated using weighted sum of legendre polynomials. Prosodic feature set includes weights of first four legendre polynomials (w_{0k}, w_{1k}, w_{2k}, w_{3k}), average jitter, average shimmer, voiced duration, total duration and change in energy of each intonation phrase. The effectiveness of the technique is demonstrated using a text dependent database of mimicked speeches. Evaluation is done by dynamic time warping of prosodic features derived from the mimicked speech and the original speech. The scores obtained from this evaluation is compared with the results of manual perception/listening tests, which clearly indicate the effectiveness of the proposed technique.

Index Terms— Prosody, intonation, mimicked speech, legendre coefficients, dynamic time warping

1. INTRODUCTION

Voice disguise is considered as a deliberate action of a person to falsify or conceal his identity [1]. Mimicking is one form of disguise where a person modifies his voice to sound like another person. Mimicry has evolved as a popular form of art when voices of celebrities like film stars, politicians are imitated in stage programs.

Intra–speaker variability is a unique feature of human speech. A person can modify the position of the articulators like lips, tongue, speech duration, pitch values, tempo, stress placement and loudness in order to mimic another person. According to Zetterholm, the principle of mimicking is based on the imitation of some specific characteristics of the original voice related to prosody, pitch register, voice quality and speech style [2] [3]. In theory, a subject could modify prosodic parameters such as intonation, loudness and rhythm by changing pitch, intensity and duration to mimic speaking style of another person. At the production level, these changes are brought out by varying the vocal fold tension, subglottal pressure and airflow, to change pitch, intensity and duration respectively [2] [4].

When a mimicry artist imitates a celebrity, the listener has certain expectations about certain features. The quality of mimicking is evaluated based on the features expected. An artist has to imitate these features to convince the listeners. Focusing on important features as well as exaggerating characteristic features may be a conscious way improving the quality of mimicking. However, there are certain minute differences between speakers, which cannot be changed, making it hard to produce exact replica of another person's voice and speech [2].

The widespread use of telephones has resulted in an increased use of human voice as an instrument in the commission of crimes [5]. In most cases, criminals who make a terror claim or a miscellaneous call, disguise their voices to hide their identity or to take the identity of another person [6]. In the context of hoax calls and impersonation, automatic systems can do voice matching. Earlier research attempts in this direction focus mostly on spectral features represented using Mel-Frequency Cepstral Coefficients (MFCC). Zetterholm analysed mimicked speech produced by one professional impersonator in terms of vowel formants, mean F_0, vowel durations and articulation rate [7] [8]. But the dynamics of pitch, energy and duration, which seemed important to human listeners, are not explicitly used by the automatic voice matching systems. In this work, our focus is on the use of prosodic features derived from pitch contour, duration and energy, for automatic evaluation of mimicked speech for the text dependant cases.

Remaining part of the paper is organized as follows. Sect. 2 describes the database used for evaluation of mimicked speech. The prosodic characteristics of original and mimicked speeches are analyzed in Sect. 3. The method used for automatic extraction of these features is discussed in Sect. 4. In Sect. 5, the mimicked speech evaluation is described along with experimental results. Sect. 6 summarizes the observations of the study.

2. DATABASE

Mimicked speech was collected from fifteen well-known professional mimicry artists. Professional mimicry artists were chosen because of the flexibility in their imitations than the amateur, concerning voice quality, intonation and articulation rate [9]. The database used for the study consists of two different parts, namely the text dependent part and the text independent part. For text dependent part, sentences were chosen from popular movies/interviews of the celebrities, and artists were asked to imitate the same sentence. For the text independent part, artists were given flexibility to speak any text of their choice while imitating celebrities. Celebrity speeches were downloaded from Internet/TV channels, whereas the mimicked versions were collected in a laboratory environment.

3. PROSODIC FEATURES FOR EVALUATION OF MIMICKED SPEECH

Characteristics such as rhythm, timing and stress lend naturalness to speech and they are collectively referred to as prosody. Prosodic cues include stress, rhythm and intonation. These include variation in pitch, relative intensity of pronunciation of sound units, correlation of speech segments according to length, overall speech tempo and pauses. Each cue is a complex perceptual entity, expressed primarily using three acoustic parameters: pitch, energy and duration.

Pitch is a perceptual attribute of sound. The physical correlate of pitch is the fundamental frequency (F_0) of vibration of vocal folds. Fundamental frequency reflects speaker-specific characteristics due to the differences in physical structure of the vocal folds among speakers. Variation of pitch as a function of time is called intonation, and is represented by the F_0 contour. It has been noticed that there are differences in intonation and duration characteristics across mimicry artists while imitating same celebrity even while uttering the same text. Fig.1 illustrates this, with speech waveform and corresponding F_0 contour side by side for a text dependent case in the database.

From Fig. 1, it can be observed that the sentences are actually uttered as different intonation units, with long pauses separating them [10]. We refer to them as 'intonation phrase'. This may not resemble the definition of 'phrase ' in a language.

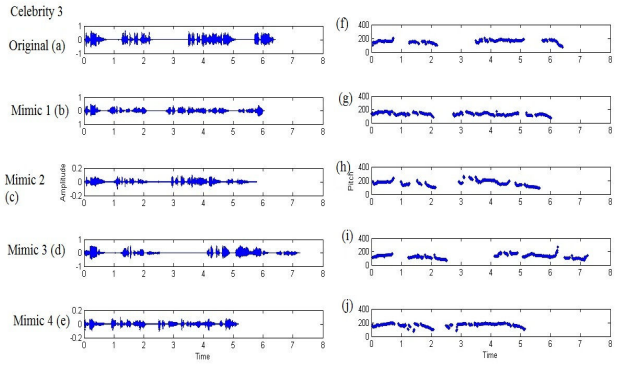

Fig. 1. Speech waveform of (a) celebrity 3 and (b), (c), (d) and (e) four professional mimicry artists mimicking the celebrity and (f), (g), (h), (i), (j) corresponding variations in F_0 contour while uttering the same text

Breaking of sentences into intonation phrases may be also due to the varying lung capacity of a person, which may influence the style of speech. The intonation and duration for each intonation phrase is different for mimicked speeches compared to original, and hence may be a useful cue for evaluation of mimicked speech.

4. EXTRACTION OF PROSODIC FEATURES

The given input speech is segmented as intonation phrases by identifying long pauses separating them. For this, speech/non-speech classification is done for each frame (20msec) independently based on three features, namely (i) short time energy (ii) the most dominant frequency (MDF) in the spectrum (iii) voicing information.

Short time energy (STE) E_n is computed using the expression:

$$E_n = \sum_{m=1}^{L} (x[m]w[n-m])^2 = \sum_{m=1}^{L} x^2[m]w^2[n-m] \quad (1)$$

where L is the number of samples of the speech signal, $w[n-m]$ represents a time shifted window sequence, whose purpose is to select a segment of the sequence $x[m]$ in the neighbourhood of sample $m=n$. The most dominant frequency (MDF) is the maximum value of the spectrum magnitude. Fig. 2 (b) and (c) shows the plot of the short term energy and most dominant frequency corresponding to a speech waveform. Voicing information shown in 2 (d) has a binary value, which indicates the presence/absence of periodicity in a frame. Three independent speech/nonspeech decisions are made for each frame based on these three features and the final decision is made using majority-voting method.

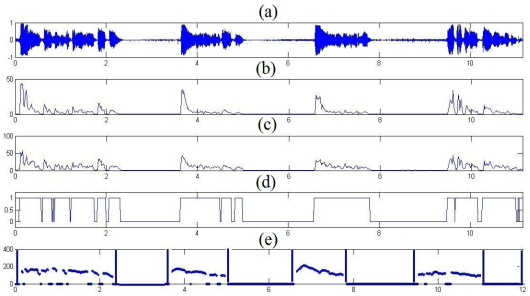

Fig. 2. Features for Speech /Nonspeech Decision a) wave form b) short time energy c) most dominant frequency d) voice information e) F_0 contour with segmented boundaries

4.1. Representation of intonation

Method used in this work for the automatic extraction of intonation features is given in Fig. 3.

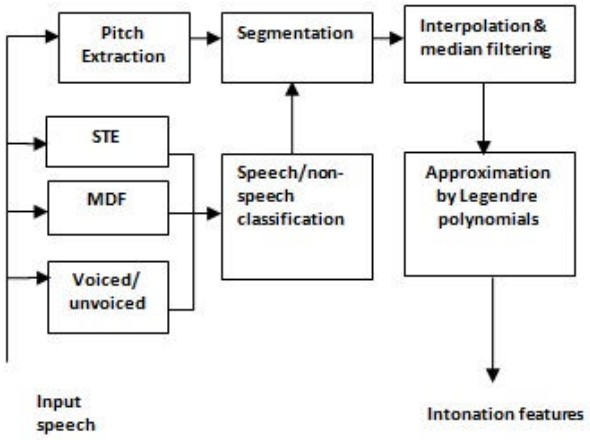

Fig. 3. Block diagram showing the extraction of intonation features

The pitch contour corresponding to each intonation phrase is interpolated, median filtered and then approximated by an M^{th} order Legendre polynomial in the sense of minimum mean square error [10].

$$f_k = \sum_{i=0}^{M} w_{ik} P_i \qquad (2)$$

where k is the pitch contour index, M is the highest polynomial order, P_i is i^{th} order Legendre polynomial and w_{ik} is weighting factor for P_i . In most cases, small value of M is sufficient and hence we have chosen M=3 for this work. Here P_0, P_1, P_2, P_3 stand for the height, the slope, the curvature, and the S curvature of F_0 contour, respectively [11]. Weighting factors of polynomials w_{0k}, w_{1k}, w_{2k}, and w_{3k} are used for representing the intonation. From Fig. 1, we can observe that the fluctuations in fundamental frequency also convey some speaker specific information and jitter is a useful parameter to indicate this. Jitter is a measure of period-to-period fluctuations in fundamental frequency, and is calculated between consecutive voiced periods via the formula:

$$\text{Jitter} = \frac{|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \qquad (3)$$

where T_i and T_{i+1} refers to the pitch period of the i^{th} and $(i+1)^{th}$ window respectively. N is the total number of voiced frames in the utterance. In this work, average jitter for an intonation phrase is used as a feature along with the weights of the polynomials.

4.2 Representation of duration

Duration characteristics of intonation phrase are represented by total duration and voiced duration that are indicative of speaker characteristics [12].

4.3 Representation of energy

Since absolute values of energy/intensity is also controlled by the settings while recording the speech, change in energy ΔE (difference between maximum and minimum energy) for an intonation phrase is included in the feature set [10]. In order to represent fluctuations in the energy, shimmer is also included in the feature set. Shimmer is a parametric measure of the instability of amplitude in each cycle, and is calculated as a measure of the period-to-period variability of the amplitude value, expressed as:

$$\text{Shimmer} = \frac{|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \qquad (4)$$

where A_i and A_{i+1} is the peak amplitude value of the i^{th} and $(i+1)^{th}$ window respectively. N is the number of voiced frames. As in the case of jitter, shimmer for a phrase is used in this case. List of prosodic features used for this study are given in Table 1.

5. EVALUATION OF MIMICKED SPEECH

Original as well as mimicked speeches for celebrity 3 (case illustrated in Fig. 1) are segmented at fairly long non-speech regions. Pitch contour along with segmented intonation phrase boundaries are illustrated in Fig. 4.

Table 1. List of prosodic features used in the study for evaluation of mimicked speech

Sl. No.	Category	Features
1.	Intonation	w_{0k}, w_{1k}, w_{2k}, w_{3k}, average
2.	Duration	total duration, voiced duration
3.	Energy	ΔE, average shimmer

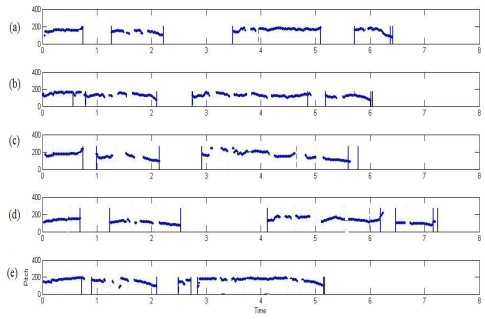

Fig. 4. Segmented pitch contour for (a) celebrity 3 and (b), (c), (d) and (e) four professional mimicry artists while mimicking the celebrity (text dependent case).

The sequences of 9-dimensional prosodic feature vectors extracted from original and mimicked speech are compared using Dynamic Time Warping (DTW) algorithm. DTW is used for measuring similarity between two sequences of vectors, which may vary in time or speed. It calculates an optimal warping path between two sequences. This warping length gives a measure of similarity between original and mimicked speech. To compare with the scores of human perception, warping length d is converted to a score using e^{-md} where m refers to a constant scaling factor. A perception/listening test was carried out with the help of fifteen human listeners. Listeners were asked to grade each mimicked utterance by choosing one among the six opinion grades, and these grades were later converted to a numerical score as; very good (11), good (8), satisfactory (6), bad (4), very bad (1). Mean Opinion Score (MOS) for each mimicked utterance was computed by taking average of the scores given by all the fifteen listeners. The mimicked speech that gets the maximum score is identified as the best one for that celebrity. Scores of automatic prosodic evaluation and perception test are given in Table 2. The entries in bold face show that the automatic evaluation results match with that of perception tests, in all the cases.

6. SUMMARY

The evaluation of mimicked speech in text dependent mode using prosodic features was performed. For extracting prosodic features, speech is segmented into phrase-like regions. Features are derived to represent duration, dynamics of F_0 contour and energy variations. Evaluation of a particular mimicked speech is done using these features, by performing DTW with the features from the original speech. The best-mimicked attempt is identified as the one with highest score. This matches with the perception test in all cases, indicating the effectiveness of proposed technique.

Table 2. Results of mimicked speech evaluation by automatic prosody matching and perception test.

Evaluation of mimicked speech for	Score of	
	Prosodic matching	Perception test
Celebrity 1	**0.65**	**5.55**
	0.62	3.5
	0.60	4
	0.55	2.5
	0.57	3.1
	0.53	3
Celebrity 2	0.28	6.4
	0.27	6.5
	0.46	**7.5**
	0.30	4.5
Celebrity 3	**0.76**	**6.1**
	0.72	6.1
	0.66	6
	0.71	4.9
	0.70	5.3
Celebrity 4	0.61	6.7
	0.68	**7.5**
	0.63	4.5
	0.55	4.9
	0.60	5.2
Celebrity 5	0.45	6.1
	0.56	**7.4**
	0.43	4.7
	0.49	5.7
Celebrity 6	0.34	5
	0.38	5.7
	0.38	6.6
	0.44	**7.1**
	0.35	3.9
Celebrity 7	**0.61**	**7.2**
	0.61	5.1
	0.52	5.5
	0.51	5.5
	0.57	5.7

Acknowledgement

Authors would like to thank Kerala State Council for Science, Technology and Environment, India for the financial support to carry out the work described in this paper.

7. REFERENCES

[1] P. Perrot, G. Aversano, and G. Chollet, "Voice disguise and automatic detection: review and perspectives," *in Lecture notes in computer science: Progress in nonlinear speech processing*, Berlin: Springer, Vol. 4391, pp. 101–117, 2007.

[2] E. Zetterholm, "Same speaker–different voices. A study of one impersonator and some of his different imitations," in *Proceedings of the 11th Australian International Conference on speech science and Technology*, pp. 70-75, 2006.

[3] E. Zetterholm, KPH Sullivan, "The impact of semantic expectation on the acceptance of a voice imitation," in *Proceedings of the 9th Ausralian conference on speech science and technology*, pp. 291-296, 2002.

[4] G. Ramya, S.A. Thati, K. Venkat and B. Yegnanarayana, "Analysis of mimicry speech", *Proceedings of Interspeech*, Portland, USA, pp.3141-3152, 2012.

[5] A. Drygajlo, "Forensic automatic speaker recognition," *IEEE Signal processing Magazine,* pp.132-135, 2007.

[6] P. Perrot, and G. Chollet, "The question of disguised voice," *Proceedings of Acoustics08*, Paris. pp.5681-5685, 2008.

[7] E. Zetterholm E., M. Blomberg, D. A. Elenius, "Comparison between human perception and a speaker verification system score of a voice imitation," *Proceedings of the 10th Australian International Conference on speech science and Technology*, pp. 393-397, 2004.

[8] E Zetterholm, "Same speaker–different voices. A study of one impersonator and some of his different imitations," *Proceedings of the 11th Australian International Conference on speech science and Technology*, pp. 70-75, 2006.

[9] M. Farrús, M. Wagner, D. Erro, and J. Hernando, "Automatic speaker recognition as a measurement of voice imitation and conversion" *The International Journal of Speech, Language and the Law*, ISSN 1748-8885. Vol. 17, no.1, pp. 119–142, 2010.

[10] L. Mary, K. K. Anish Babu, and A. Joseph, "Analysis and Detection of mimicked speech based on prosodic features," *International Journal of Speech Technology,* Vol. 15, pp. 407–417, 2012.

[11] C. Lin and H. Wang, "Language identification using pitch contour information," *Proc. Proceedings of Int. Conf. Acoust., Speech and Signal Processing,* vol. I, pp. 601-605, 2005.

[12] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, pp. 782-796, 2008.