ANALYSIS OF MIMICRY SPEECH

D. Gomathi, Sathya Adithya Thati, Karthik Venkat Sridaran and B. Yegnanarayana

Speech and Vision Lab, International Institute of Information Technology, Hyderabad, India

{gomathi, sathya.adithya, karthik.venkat}@research.iiit.ac.in and yegna@iiit.ac.in

Abstract

In this paper, mimicry speech is analysed using features at suprasegmental, segmental and subsegmental levels. The possibility of the imitator getting close at each of these levels is examined here. The imitator cannot duplicate all features of the target, as imitation depends on the target speaker, utterance chosen, and his ability to imitate. To study the variation of features in the case of best and poor imitations, the source and system features are observed for different target speakers and for different utterances. Features such as pitch contour, duration, Itakura distance, strength of excitation and loudness measure are used for this analysis. Perceptual evaluation is performed to determine the closeness of imitation to the target. The closeness of features for best imitated and poorly imitated utterances is presented here.

Index Terms: Mimicry, imitation, suprasegmental, segmental, and subsegmental features

1. Introduction

Speech is a major form of communication used by human beings. It conveys the mood of the speaker by variations in pitch, loudness, intonation, stress, pause and other such features, due to flexibility of human speech production mechanism. Mimicry/voice imitation is a fine art, where the mimicry artist trains his voice to imitate the voice of a target speaker. The ability to produce sounds close to the target improves with practice. The imitator's voice gets close to the target's voice depending on the way he controls his speech production system.

Since the speech production system is different for each person, it is difficult for the imitator to sound exactly the same as target speaker. Usually for entertainment purposes, the imitator tries to sound like a caricature of the target by exaggerating some prominent features[1].

Imitating a target speaker involves capturing the dialectal variations, accent, speaking style, pronunciation, and intonation of the target speaker. Imitation also involves mimicking some, if not all, of the following: body language, nonverbal cues, gestures, and typical phrases. The imitated features may be at suprasegmental, segmental and subsegmental levels. The imitator makes lot of effort in imitating another person, as it is not natural for him to modify all the above mentioned features at the same time. The terms mimicry and imitation are used interchangeably in this paper.

Analysis on imitation/mimicry speech is limited, as there is no standard database available. Also, it is difficult to find a professional imitator whose mimicry speech is close to the target speech. Analysis of imitation from auditory, acoustic and phonetic perspectives has been carried out by Zetterholm [1]. The features of a good impostor speech that generally match the target speech are duration, mean fundamental frequency (F0), articulation rate and formant frequencies [2]. It is also reported that an imitator can imitate the global timing of the target, but not the local timing. The studies in [2] reveal that durations of individual words are closer to durations of imitator's natural voice than the target. In [3], Zetterholm has reported that the imitator is able to change duration at word level and it is not close to the durations of his natural voice. Discrete Fourier Transform (DFT) spectra and the difference between the amplitudes of the first and second harmonics (H1-H2) have been analyzed in [4]. Estimation of cross sectional area of tube sections reflect the tendency of the imitator to change the cross sectional areas of the vocal tract specific to the target speaker [5]. In [6], Endres et al have concluded that an imitator can change the formant positions of his voice within certain limits. It is also reported that the formant structure of imitator and target do not agree, especially in high frequency bands. In this paper mimicry speech is analysed at various levels to see which features tend to move towards the target feature space for best and poor cases of imitation. Perceptual evaluation is conducted to choose the best and poorly imitated utterances for analysis.

The terminology used in the paper is similar to the one used in [5]. The utterance spoken by the Indian celebrity (actor) will be referred to as target [T]. The utterance spoken by the imitator, when he imitates the target (celebrity), will be referred to as imitation [I]. The utterance spoken by the imitator in his original voice will be referred to as natural [N].

The paper is organised as follows. In Section 2, the data collection procedure is explained. Perceptual evaluation method and its results are discussed in Section 3. Section 4 discusses the features used for analyzing the mimicry speech. The results and conclusions are presented in Section 5.

2. Data collection

The main challenge involved in performing analysis of mimicry speech is data collection, due to lack of sufficient good data. Data for the analysis of mimicked speech was recorded by a professional mimicry artist, who has been practicing the art for the past 15 years. Data was collected at a sampling frequency of 48 kHz in a recording studio (clean environment). Recordings of five popular Indian celebrities voices (PO, MB, PR, NG, SP) were collected from interviews and movies. The utterances chosen were from Telugu language which is a regional language in the southern part of India. Ten utterances for each target [T] were chosen. The duration of each utterance varies from 2 to 12 seconds. Utterances of short duration do not contain many prominent prosodic features, and the imitator has to be very good to imitate such utterances. All the target utterances were imitated by the professional imitator five times. Recording of the utterances was done in his natural voice [N] as well.

3. Perceptual evaluation

3.1. Evaluation 1

Subjective evaluation was conducted using 10 listeners to evaluate the quality of the mimicry data. The listeners were native speakers, and have knowledge of the target's voice. All the listeners were presented with utterances in the target voice and five repetitions of imitated utterance by the imitator. They were asked to score the similarity of the imitated speech utterance to the target utterance on a scale of 1 to 5 (1: Highly dissimilar, 2: dissimilar, 3: somewhat similar and somewhat dissimilar, 4: similar, 5: Highly similar). The evaluation scores are presented in Table 1. Results (Mean scores from all the listeners for all utterances) indicate that the mimicry artist has imitated PO (celebrity) and MB (celebrity) well. Scores also indicate that the quality of mimicry speech is good. The best and poorly imitated utterances chosen for analysis are utterance 1 of celebrity PO and utterance 4 of celebrity SP respectively.

Table 1: *Mean scores of subjective evaluation*

Utterance number of the target	PO	MB	PR	NG	SP
1	4.27	3.63	3.27	3.18	4.09
2	3.63	4	2.5	3.63	2.95
3	3.73	3.63	3.09	3.95	3.54
4	3.5	3.36	2.32	2.63	2.3
5	3.0	3.36	3.18	2.95	3.68
6	3.64	3.45	2.77	3.45	3.54
7	3.59	3.73	3.36	3.73	3.77
8	3.27	4.19	2.81	3.59	3.27
9	4	3.77	2.36	2.95	3.68
10	3.76	3.55	3.09	3.05	3.27
Average of all utterances	3.64	3.67	2.88	3.31	3.41

3.2. Evaluation 2

A second type of blind evaluation was performed by a set of 30 listeners who did not participate in the first evaluation. All the listeners were presented with the imitated utterances of all targets, and the task was to identify the target (famous celebrity), and tell if it is an original or imitated utterance. Here the listeners have the knowledge of target (celebrity) voice. The files were resampled to 8 kHz, as human listeners can identify the target even over telephone channel. The experiment was conducted to see if listeners were able to identify the target speakers from the imitated utterance. Out of 30 imitated utterances, for 21 utterances the targets were identified correctly and 16 of them were reported as spoken by celebrity (original). This evaluation confirmed that the imitator was good at imitating the targets most of the time.

4. Features for analysis

Speech signal can be analysed at three different levels, namely, subsegmental, segmental and suprasegmental, based on the size of the segment used for analysis. The subsegmental features are extracted over a very short (1-5 ms) analysis window, typically less than a pitch period. The subsegmental features used for analysis are strength of excitation (SoE) and perceived loudness measure [7, 8].

The segmental features are extracted over a short (10 - 30 ms) interval of time, during which the signal is assumed to be stationary. Most of the time speech signal is analysed using segmental features like spectral features, which represent the characteristics of the vocal tract shape. In this work, the linear prediction coeffcients are used to represent the segmental features.

Suprasegmental features mainly refer to the behavioral aspects (speaking habits) of a speaker, and are typically extracted over a large ($>$ 200 ms) analysis window. Intonation (pitch contour), syllable durations and speaking rate are some of the suprasegmental features. These are the features which human beings tend to use for imitation. It is likely that these are the features that dominate in perception.

4.1. Suprasegmental features

4.1.1. Pitch contour

An imitator tries to change his F0 contour so that the shape of the contour matches with the target F0 contour, i.e., the rising and falling of the F0 values in the contour are as close as possible. The F0 values in the contour are extracted using Zero Frequency Filtering on the speech signal [9]. The method involves passing the differenced speech signal twice through a digital resonator having poles at zero frequency. The trend in the output is removed by local mean subtraction using a window length in the range of 1 or 2 pitch periods. The negative to positive zero-crossings in the zero frequency filtered (ZFF) output give the glottal closure instants or epochs . The reciprocal of the interval between two successive epochs gives us the instantaneous fundamental frequency.

The average F0 of the target, imitation and natural are compared in Table 2. From the table we observe that the imitator is able to imitate both increased and decreased average F0 of the target in most of the cases, except for the case of NG. The contours of the instantaneous fundamental frequency (F0) after time alignment of I *vs* T and I *vs* N are plotted in Figure 1 for best imitated utterance, and in Figure 2 for poorly imitated utterance. We see a good match of the pitch contours in the case of best imitation, while there is poor match in the case of poorly imitated utterance.

A deviation measure is used to compute the deviation in the pitch contours. All $F0$ values are normalized by dividing them with the mean $F0$ value. The sum of squared difference of these normalized values has been computed. The deviation scores are presented in Table 3. We see from the Table that in the best cases of imitation, the deviation between I *vs* T is less, whereas the deviation between I *vs* N is high.

Table 2: *Mean F0 values (in Hz) for Target, Imitation and Natural voices for best imitated utterance of each celebrity*

Celebrity	Target	Imitation	Natural
PO	278.8485	278.0185	148.6121
MB	188.9273	183.3305	124.2880
PR	117.5183	120.3016	166.4184
NG	149.4316	132.3516	137.2193
SP	118.1604	111.3066	124.0396

4.1.2. Duration

The imitator tries to capture the global duration characteristics of the target. He pauses and hesitates at the same instants as the target. When the target speaker is silent for some duration, the imitator also pauses, but the durations of silence need not match

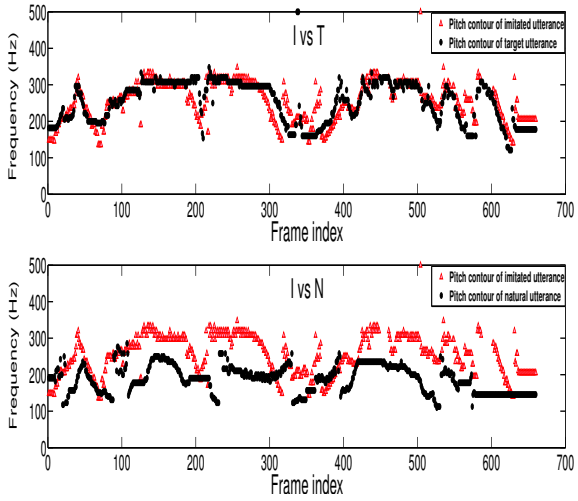

Figure 1: *Pitch contour of Target, Imitation and Natural voice of best imitated utterance*

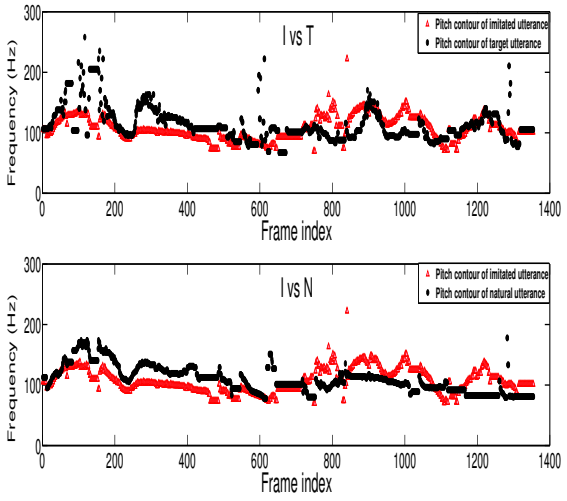

Figure 2: *Pitch contour of Target, Imitation and Natural voice of poorly imitated utterance*

Table 3: *Deviation of F0 values for Target, Imitation and Natural voices for best imitated utterance of each celebrity*

Celebrity	I *vs* T	I *vs* N
PO	64.65	89.88
MB	29.83	38.98
PR	41.21	52.14
NG	21.04	14.16
SP	99.77	81.01

well. The mean global duration values for all utterances of each target are shown in Table 4.

Table 4: *Mean duration values (in sec.) for Target, Imitation and Natural voices of all utterances*

Celebrity	Target	Imitation	Natural
PO	4.176	3.896	3.692
MB	2.495	2.489	2.043
PR	2.67	2.468	2.474
NG	3.074	2.905	2.439
SP	5.508	5.571	5.293

4.2. Segmental features

These are features extracted using an analysis window of 20 ms duration. The utterances were time aligned using Dynamic Time Warping. Linear Prediction Coefficients (LPC) were extracted from speech signal for every 20 ms with a frame shift of 5 ms. Itakura distance was computed between I *vs* T and also between I *vs* N. The distances show that the imitator is close to his original voice than the target voice. The mean Itakura distance between I *vs* T and I *vs* N for the best imitated utterance of each celebrity are given in Table 5. We observe that the distance between T and I is higher than the distance between I and N.

Table 5: *Mean values of Itakura distance for Imitation vs Target and Imitation vs Natural for best imitated utterance of each celebrity*

Celebrity	Imitation *vs* Target	Imitation *vs* Natural
PO	1.0176	0.6053
MB	1.0222	0.8372
PR	1.2729	0.7678
NG	0.6812	0.5062
SP	0.7426	0.5305

4.3. Subsegmental features

While it appears to be easy to imitate suprasegmental features, it is not clear whether subsegmental features also can be imitated or not. The subsegmental features considered here are SoE and loudness measure [7, 8].

4.3.1. Strength of excitation

The strength of excitation (SoE) is related to the strength of the impulse-like excitation of the glottal activity. The SoE is derived from the slope of the ZFF signal at epoch locations [7]. The scatter plots of SoE *vs* fundamental frequency ($F0$) are shown in Figure 3 and Figure 4 (best viewed in color). We observe that the imitator tries to change SoE values to match the target for the best case as depicted by Figure 3. Figure 4 shows that most of the SoE values are lying close to his natural voice. In general, the cluster of points for the imitator in the $F0$ and SoE plane tend to move towards the cluster of points of the target, indicating the effect of importance of $F0$ and SoE on the perception of imitation.

4.3.2. Measure of loudness

Perceived loudness of speech is also related to the abruptness of the glottal closure. An objective measure (η) of perceived loudness based on the abruptness of glottal closure derived from the speech signal is discussed in [8]. When the glottal closure is abrupt, the Hilbert envelope of the LP residual of the speech signal will have sharper peaks at the epochs. The sharpness of

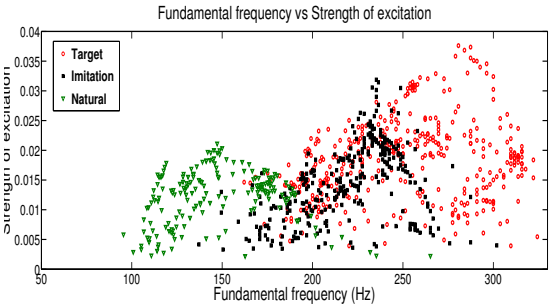

Figure 3: *Scatter plot of strength of excitation of Target, Imitation and Natural voice of best imitated utterance*

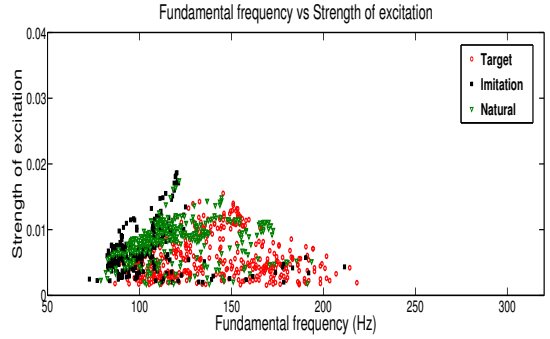

Figure 4: *Scatter plot of strength of excitation of Target, Imitation and Natural voice of poorly imitated utterance*

the peaks in the Hilbert envelope at the epochs is derived by computing the ratio $\eta = \mu/\sigma$. Here μ denotes the mean, and σ denotes the standard deviation of the samples of the Hilbert envelope of the LP residual in a short interval (2 ms) around the epochs. The mean loudness measure of all the utterances of all target speakers is given in Table 6. The tendency to move towards target space can be observed in the case of PO, MB, PR, and NG. The loudness measure is a useful feature for classifying breathy and modal voice. The η values are lower for breathy voice as compared to modal voice. In case of celebrity NG, the voice type is breathy. The imitator tries to sound breathy when he imitates NG and the values of loudness measure become lower as compared to his original voice. Table 7 compares the mean η values of T, I and N in the case of NG.

Table 6: *Mean η values for Target, Imitation and Natural voices of all utterances*

Celebrity	Target	Imitation	Natural
PO	0.634	0.628	0.686
MB	0.642	0.647	0.687
PR	0.606	0.655	0.673
NG	0.556	0.59	0.652
SP	0.624	0.718	0.707

5. Conclusions

In this paper we have analysed imitated speech along with target and natural speech using various features at suprasegmental, segmental and subsegmental levels for both good and poor imitation cases. The analysis was performed using the following features: pitch contour, duration, strength of excitation, loudness measure and Itakura distance.

Table 7: *Mean η values for Target, Imitation and Natural voices of five utterances of celebrity NG*

S.No. of utterance	Target	Imitation	Natural
1	0.5468	0.5448	0.6440
2	0.5542	0.5665	0.6628
3	0.5449	0.5988	0.6421
4	0.5511	0.5981	0.6428
5	0.5514	0.5443	0.7137

The analysis shows that features at suprasegmental and subsegmental levels are mostly varied during imitation. For well imitated cases, features at both these levels, seem to move towards the target. In some poorly imitated cases, features at either suprasegmental or subsegmental levels seem to move towards the target. It appears that movement of features at either level produces the perception of imitation.

Segmental features, which represent the vocal-tract shape of the speaker are difficult to manipulate. Itakura distance, used as a feature at segmental level, was higher between target and imitation than for imitation and natural in all cases of imitation. This shows that it may be difficult to match the spectral features at the segmental level during imitation, as the spectral characteristics are dependent on the size and shape of the vocal tract system of the individual.

6. Acknowledgements

The authors would like to thank Mr. Janardhan and Mr. Anthony Raj for high quality imitations and to the listeners who have participated in subjective evaluations.

7. References

[1] Elizabeth Zetterholm, " Same Speaker: different voices: A study of one impersonator and some of his imitations", Proc. Int. Conf. Speech Sci. and Tech., pp. 70-75, 2006.

[2] Anders Erikkson and Par Wretling, "How Flexible is the Human Voice? – A Case Study of Mimicry", Eurospeech97, Rhodes, Greece, pp. 1043 – 1046, 1997.

[3] Elizabeth Zetterholm, "Intonation pattern and duration differences in imitated speech", In Proc. Speech Prosody 2002, pp. 731 – 734, Aix-en-Provence, 2002.

[4] Tatsuya Kitamura, "Acoustic Analysis of Imitated Voice Produced by a Professional Impersonator", Interspeech, pp. 813 – 816, September 2008.

[5] Gal Ashour and Isak Gath, "Characterization of Speech during Imitation", Eurospeech99, Budapest, Hungary, September 1999.

[6] W. Endres, W. Bambach, and G. Flosser, "Voice spectrograms as a function of age, voice disguise and voice imitation",Journal of the Acoustical Society of America, 49:1842-1848. 1971.

[7] K.S.R. Murthy, B.Yegnanarayana, and M.Anand Joseph, "Characterisation of glottal activity from speech signals", IEEE Signal Processing Letters, vol.16, no. 6, Jun 2009.

[8] G.Seshadri and B.Yegnanarayana, "Perceived Loudness of speech based on the characteristics of excitation source", Journal of the Acoustical Society of America, Vol. 126, No.4, pp. 2061-2071, Oct 2009.

[9] K.S.R. Murthy and B.Yegnanarayana, "Epoch Extraction from speech signals", IEEE Trans. Audio,Speech, Lang Process., vol.16, no. 8, pp. 1602 – 1613, Nov 2008.