Understanding the significance of different components of mimicry speech

D. Gomathi[1]*, P. Gangamohan*[2] *and B. Yegnanarayana*[3]

Speech and Vision Laboratory, International Institute of Information Technology,
Gachibowli, Hyderabad, India 500032

gomathi@research.iiit.ac.in[1], gangamohan.p@students.iiit.ac.in[2], yegna@iiit.ac.in[3]

Abstract

Voice conversion systems aim at finding a transformation function using statistical models. Mimicry (voice imitation) is a natural voice transformation technique which sounds convincing to the listeners. It thus seems advisable to study the transformation used by human beings who perform mimicry. The objective of this study is to examine the various components of speech that are modified during voice imitation. To transform a given speech utterance to sound like that of a target utterance, the process needs to be understood at both production and perception level. In this paper, the importance of source and system parameters and also the significance of different components of speech that contribute to the perception of imitation are studied. A flexible analysis-synthesis tool is used to modify the features of natural utterance and convert it to imitated utterance. Perceptual studies are carried out to understand if the modified features contribute to imitation. The results show that a combination of features is varied by the imitator to achieve imitation and they vary depending on the target speaker.

Index Terms: Mimicry, voice imitation, speech synthesis, speech prosody.

1. Introduction

Speech is a natural medium of communication among human beings. Speech signal carries information about the intended message, the identity of the speaker and the background. There are a few features in speech signal that enable us to differentiate between speakers. There are applications in gaming industry which require voices of celebrities for gaming avatars. Voice Conversion (VC) is a technique to transform an utterance of a source speaker so that it is perceived as if spoken by a specific target speaker. A variety of techniques have been proposed for the conversion function [1], such as artificial neural networks, dynamic frequency warping or Gaussian mixture model.

Human beings imitate voice for language acquisition, entertainment or for voice disguise [2]. Voice imitation is primarily used for entertainment, where the mimicry artist trains his voice to imitate the voice of a target speaker. An imitator cannot imitate all the features of the target speaker but tries to imitate features that are perceptually significant.

In the literature, analysis of voice imitation has been carried out and the closeness of features like fundamental frequency, duration, discrete Fourier transform spectra (DFT) and formant frequencies have been studied [3]. Different studies of professional imitators and their imitations suggest the possibility of getting close to target voice. Both perceptual and acoustic analysis confirm the flexibility of human voice [2]. It is also to be noted that an imitator might exaggerate a few important features while he may ignore a few less important features [4].

Mimicry is a natural voice transformation technique which takes into account the ignored aspects of speech synthesis research. Studying the way a professional imitator imitates will help in building better voice conversion systems. For voice transformation/conversion, the cues underlying a particular voice quality need to be identified and represented. It is not sufficient to just represent them; but they need to be modified in such a way that modified speech signal sounds natural. It is more likely that more than one feature is modified during imitation and these features vary depending on the target speaker. It is also possible that two different imitators may choose different features for the same target speaker. In this paper, the components of speech that are modified by an imitator are addressed by signal processing techniques. In this study, the speech data from a professional imitator who performed mimicry of various celebrities has been used [5]. The first study addresses the importance of source and system parameters in voice imitation. The second study examines the features modified by an imitator, few experiments are carried out to synthetically transform the natural utterance spoken by an imitator to the corresponding imitated utterance by varying different features of speech. This is carried out using a flexible analysis-synthesis tool (FAST) [6]. This tool was used to transform a neutral utterance to an emotional utterance and vice versa. In FAST, two utterances spoken by the same speaker are matched using dynamic time warping (DTW) algorithm to get two warping paths. These warping paths are used for understanding the way different features of speech are modified. The features correspond to both source and system characteristics of speech production mechanism.

The terminology used in the paper is similar to the one used in [7]. 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 terms mimicry and imitation are used interchangeably in this paper.

The paper is organized as follows: Section 2 discusses the data and the features of speech that will be used for modification of natural utterance. In Section 3, the following experiments are conducted: (i) to understand the significance of source and system parameters in imitation and (ii) to convert a natural utterance to an imitated utterance. The results of subjective studies on the experiments are also discussed. Section 4 gives the summary and conclusion.

2. Data and Feature Extraction

Database for the current study consists of recordings by a professional mimicry artist in Telugu language [5]. Voices of five popular Indian celebrities (MB, NG, PO, PR and SP) were chosen as target. These voices were collected from interviews and

10.21437/SpeechProsody.2014-201

movies. For each target voice, ten utterances of short duration were chosen. 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 as well. There are three parallel utterances corresponding to target (T), imitation (I) and natural utterance (N) in the database.

The source filter theory of speech states that speech can be described as the output of sound source being modulated by a dynamically varying filter. The speech signal carries information about the dynamic vocal tract system and the excitation source. The vocal tract system and excitation source parameters represent inherent characteristics of the speech signal, while duration and intonation are examples of acquired characteristics over a period of time. The features that are extracted for this study are vocal tract features, excitation source features and prosodic features.

The linear prediction coefficients (LPCs) are used to represent the vocal tract information. They are obtained using LP analysis method. The basic idea is that a speech sample at time instant n, can be approximated as a linear combination of the past p, speech samples.

$$\tilde{s}[n] = \sum_{k=1}^{p} a_k s[n-k] \tag{1}$$

where s[n] are speech samples, $\{a_k\}$ are the predictor coefficients and $\tilde{s}[n]$ are the predicted samples.

An all-pole filter H(z) is used to represent the vocal tract parameters of the speech signal in the frequency domain.

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

The error in prediction is given by

$$e[n] = s[n] - \tilde{s}[n] \tag{3}$$

The representation in frequency domain is given as

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{4}$$

As A(z) is the reciprocal of H(z), LP residual is obtained by the inverse filtering of speech.

The LP residual obtained from LP analysis is used for the representation of excitation source information. The successive samples in the LP residual are less correlated compared to the samples in the speech signal.

Prosodic features are represented by the pitch contour and duration parameter. The pitch contour which is modified by the imitator is derived using zero frequency filtering (ZFF) method [8]. 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 one to two pitch periods. The negative to positive zero-crossings in the zero frequency filtered output give the glottal closure instants or epochs. The reciprocal of the interval between two successive epochs gives the instantaneous fundamental frequency.

Duration parameter of natural and imitated utterances are mapped frame-wise using dynamic time warping (DTW) algorithm. DTW is an algorithm used for measuring optimal match between two utterances which may vary in time or speed. The utterances are represented by a sequence of vectors which correspond to the vocal tract features. The DTW algorithm is constrained as the labelling of data is done phoneme-wise.

3. Understanding the relative importance of components of speech in imitation

3.1. Significance of source and system parameters

During imitation, the imitator tries to position his articulators in some specific way in order to imitate a few target speakers. Though there are physiological constraints on the vocal tract, there is some flexibility in positioning tongue and some articulators. This brings the changes in his system (vocal tract) parameters. The imitator has to modify the way he excites his vocal folds to produce some of the voice characteristics. This brings the changes in the excitation source characteristics.

A study was performed to understand the importance of source and system parameters in performing voice imitation. A 10^{th} order short-term (20 ms frame size and 10 ms frame shift) LP analysis is performed to compute the residual signal and LP coefficients. The LP coefficients are converted to 20 dimensional linear prediction cepstral coefficients (LPCCs). The LPCCs and residual are extracted for 'T','I' and 'N'. The speech signals are time aligned using dynamic time warping (DTW) with LPCCs as feature vectors. For synthesis, the residual of the imitated utterance is passed through LP filter corresponding to the system parameters of the natural utterance. All combinations of residual and LP coefficients of 'T', 'I' and 'N' of all celebrities (MB, NG, PO, PR, SP) were used for synthesis to know the importance of source and system parameters.

Table 1: *Subjective evaluation results for all combinations of source and system parameters of 'T', 'I' and 'N'.*

Experiment	Source	System	MB	NG	PO	PR	SP
E1	I	T	1	1	1	1	1
E2	T	I	1	1	1	1	1
E3	N	T	1	0	0	0	1
E4	T	N	0	1	1	1	0
E5	N	I	1	0	0	0	0
E6	I	N	0	1	1	0	0

The synthesized files obtained after interchanging the corresponding source and system features for all cases mentioned above are assessed by subjective evaluation. The evaluation is carried out by twenty listeners in the age group of 21-30. Each subject was given six synthesized files and asked to give a score of '1' if it is target (T)/ imitation (I), '0' if it is natural (N). The results of the evaluation are presented in Table 1. The scores in the table are arrived by majority voting. All the synthesized speech files were presented in random order, and were not grouped in any particular order.

The rows E1 and E2 show that when source parameters belong to 'I' and the system parameters belong to 'T' or vice versa, the synthesized speech sounds similar to target for all celebrities.

For celebrity MB, when system parameters of 'T' or 'I' are used, the synthesized file sounds closer to 'T' as seen from rows E3 and E5. When system parameters of 'N' are used, the synthesized file sounds like an unknown speaker. The listeners reported that the characteristic pause of 'T' was missing hence the synthesized speech sounds like unknown speaker. So the system parameters seem to play a bigger role in this case.

In the case of celebrity 'NG', the voice quality is breathy. So whenever source parameters of 'T' or 'I' are used, even if the system parameters belong to 'N', there is breathiness in the synthesized speech which gives an impression that we are listening to 'T' or his imitation. This can be observed from rows E4 and E6.

The source parameters play an important role in the case of celebrity 'PO'. This is because there is an increase in loudness when the source features of 'T' are used. The listeners could make out the difference clearly between the experiments where source parameters of 'T' or 'I' were used. The effect of source features of 'I' is similar to 'T' in terms of intonation but the level of loudness is low. The use of source parameters of 'N' makes the synthesized file sound like 'N' or unknown speaker.

The results of celebrity 'PR' is similar to that of celebrity 'PO', except for row E6. This may be because the source parameters in imitation 'I' are not well imitated in case of celebrity 'PR'.

The imitations of celebrity 'SP' is similar to target for rows E1, E2 and E3. The synthesized files for rows E4, E5 and E6 sound like 'N' or unknown speaker. The expectation is when source or system parameters of 'T' or 'I' are used, the synthesized file should be similar to 'T' or 'I', but the files sounds like an unknown speaker. This may be because the imitations of celebrity 'SP' were not well imitated.

3.2. Perceptual significance of features

The aim of this study is to understand the features that are modified during imitation. The imitator's natural voice and imitation are compared and the differences in features are studied. The features from the imitated voice are incorporated into the natural utterance of the imitator so that imitated voice can be synthesized from natural voice. To modify the natural utterance in order to make it sounds similar to imitated utterance, a flexible analysis-synthesis tool (FAST) has been used. The main feature of FAST is that it can be used to match two utterances of same lexical content spoken by same speaker to determine the warping path (WP). After time alignment, modification of features is carried out as per the warping path. The modified features are then used to synthesize the imitated utterance. The synthesis is carried out using prosody modification program [9].

A 10^{th} order LP analysis is performed on a speech segment of 20 ms for every 10 ms. The 11 LP coefficients are converted to 20 dimensional LPCCs. Time alignment of the natural utterance with the imitated utterance is carried out using the DTW algorithm. The optimal warping path obtained by DTW represents the best mapping between the natural and imitated feature vectors. Two warping paths are obtained for each pair (natural and imitated) of utterances. Warping path 1 (WP1) corresponds to the one in which all frames of the imitated utterance are used. Usage of WP1 will automatically modify the duration. Warping path 2 (WP2) corresponds to the usage of all frames of natural utterance. Figures 1 and 2 show the warping paths WP1 and WP2. Pitch contour of the natural utterance is mapped with that of imitated utterance using the warping path 2 (WP2), as shown in Figure 3. Pitch and duration are modified using the prosody modification program [9]. In the case of LPC modification, the LPCs of each frame of natural utterance are replaced by the LPCs of imitated utterance. For LP residual modification, the residual between two epochs is replaced by an Liljencrants-Fant (LF) model estimate [10]. The set of experiments described in [11] were performed. The features modified and the warping paths used are mentioned in Table 2, '0' in the feature column

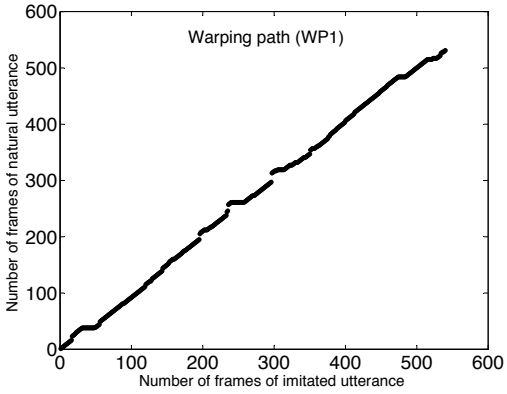

Figure 1: *Illustration of warping path (WP1) when imitated utterance is reference vector and natural utterance is test vector.*

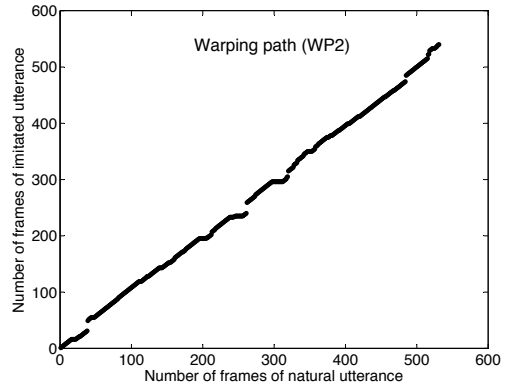

Figure 2: *Illustration of warping path (WP2) when natural utterance is reference vector and imitated utterance is test vector.*

indicates that the feature is not modified and '1' indicates that the corresponding feature is modified in this synthesis experiment.

Table 2: *Experiments and corresponding warping paths for modification of features of natural utterance.*

Experiment	Feature				Warping Path
	Residual	LPC	Duration	Pitch	
E1	0	0	0	1	WP2
E2	0	0	1	0	WP1
E3	0	1	0	0	WP2
E4	0	0	1	1	WP1
E5	0	1	1	0	WP1
E6	0	1	0	1	WP2
E7	0	1	1	1	WP1
E8	1	1	1	1	WP1

There are 10 imitated and 10 natural utterances for each of five celebrities in the database. For each utterance all the eight experiments listed in Table 2 are conducted. Ten listeners participated in the listening test to evaluate the synthesized speech

Table 3: Subjective evaluation results of synthesized imitated utterance.

Experiment	No. of features modified	Feature modified	MB-I	MB-T	NG-I	NG-T	PO-I	PO-T	PR-I	PR-T	SP-I	SP-T
E1	1	Pitch	2.89	1.71	3	2.71	2.85	1.71	3	2.28	2.67	2.28
E2	1	Duration	1.77	1.42	1.85	1.71	1.85	1.57	1.42	1.31	2.67	2.28
E3	1	LPC	1.97	1.32	1.85	1.14	1.85	1.28	1.71	1.71	1.83	1.28
E4	2	Pitch, Duration	2.78	2.31	2.85	2.42	2.71	2.28	3.42	2.14	2.5	2.5
E5	2	Duration,LPC	1.75	1.3	2.14	1.42	1.85	1.85	2.14	1.85	2.33	2.42
E6	2	Pitch, LPC	2.78	2.45	3.42	3.28	3.57	2.85	3.14	2.71	2.5	2.42
E7	3	Pitch, Duration, LPC	2.67	2.71	3	3	3.42	2.85	3.85	2.85	3	2.14
E8	4	Pitch, Duration, LPC, Residual	3.02	2.67	3.28	2.85	3.57	3.37	3.57	3.14	3.5	2.85

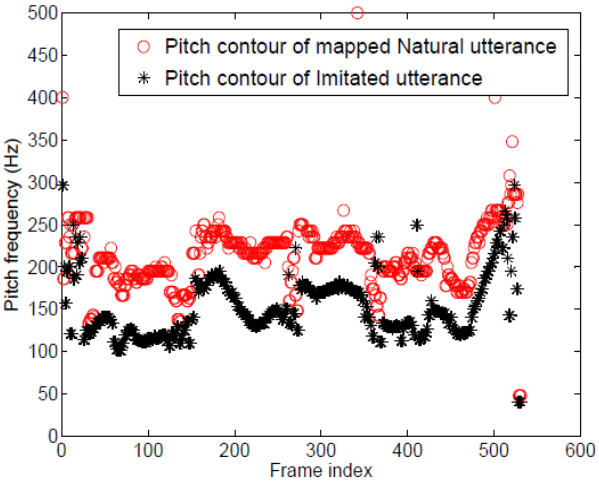

Figure 3: *Mapping of instantaneous pitch contour of natural utterance to that of imitated utterance.*

obtained after modification. The best imitated utterance of each celebrity is considered for subjective evaluation. Each subject is given a natural utterance, an imitated utterance, a target utterance and eight modified utterances which are synthesized by experiments 1 to 8 (presented in random order). The listener has to compare each synthesized utterance to imitated utterance and target utterance and give a score on a scale of 1-5 (1: highly dissimilar, 2: dissimilar, 3: somewhat similar and somewhat dissimilar, 4: similar, 5: highly similar). For example, if the synthesized utterance is compared to imitated utterance, a score of 5 indicates that synthesized file is very similar to imitated, while a score of 1 indicates synthesized file is very different from imitated. The results presented in Table 3 are the mean scores of all 10 listeners. The comparison of synthesized utterance to 'I' is performed and scores are presented in columns MB-I, NG-I, PO-I, PR-I, SP-I. Similarly the comparison between synthesized utterance to 'T' is performed and scores are presented in columns MB-T, NG-T, PO-T, PR-T, SP-T. It is expected that the scores for similarity of synthesized file to imitated utterance will be higher than the scores for similarity of synthesized file to target utterance.

The following observations are made from Table 3. The rows E1, E2 and E3 correspond to modification of one feature at a time namely pitch, duration and LPCs. The high scores in E1 indicates that pitch is a major suprasegmental feature that an imitator can modify easily and contributes more to the perception of imitation. The rows E2 and E3 show us that modification of duration and LPCs alone do not contribute as significantly as pitch modification. The rows E4, E5 and E6 correspond to modification of two features at a time. The rows E4 and E6 where pitch is modified along with duration and LPCs has better scores than E5 in which pitch is not modified. The row E4 gives lower scores for all targets except PR. This might be because duration modification is not aiding the feature pitch in perceiving imitation. The row E6 where pitch and LPCs are modified, shows significant high scores for 'NG','PO' and 'PR'. The combination of pitch, duration and LPCs seem to give significant improvement in the synthesized imitated utterance as can be seen from E7 especially for voices of 'PR' and 'SP'. E8 corresponding to the case where LP Residual is replaced by an LF model has higher scores. Though the residual is absent in this case, the perceptual scores are still high for celebrity 'MB' and 'SP'. In section 3.1, it was shown that system parameters play a big role in the imitation of celebrity 'MB', hence the absence of residual does not seem to affect the perceptual scores. The combination of pitch and LPC features give high scores for celebrities 'NG' and 'PO' but the combination of pitch, duration and LPC gives high score for celebrity 'PR'. The above results show us that the combination of features vary as per the target speaker.

4. Summary and Conclusion

In this paper, the various features of speech that contribute to the perception of imitation have been studied. The first study was to identify the contribution of source and system parameters. The subjective evaluation by listeners confirmed that source parameters were important for target speakers like 'NG' and 'PO' while system parameters were important for target 'MB'. The second study was modification of excitation source, vocal tract and prosodic features. The modification was performed using flexible analysis-synthesis tool. The synthesized files were evaluated for their closeness to imitation and target. The prosodic feature pitch contour seems to play a major role in contributing to the perception of imitation. Though duration and linear prediction coefficients individually do not contribute much to imitation but their combination along with pitch contour gives a good amount of similarity to imitation. The above observations are general ones but the combination of features also vary with the target speaker. The same combination of features need not give high perceptual scores for all target speakers. Further studies can be carried out by collecting mimicry speech from many professional artists to examine whether same features are modified by all of them for a given target speaker.

5. References

[1] Y. Stylianou, "Voice transformation: A survey", *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP)*, Taipei, Taiwan, pp. 3585-3588, April 2009.

[2] E. Zetterholm, " *Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success*", Doctoral dissertation, Travaux de linstitut de linguistique de Lund 44, Lund University,2003.

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

[4] E. Zetterholm, "Detection of speaker characteristics using voice imitation", *In C. Mller and S.Schtz (eds.) Speaker Classification, Springer LNCS/LNAI series*, 2006.

[5] D. Gomathi, Sathya Adithya Thati, Karthik Venkat Sridaran and B. Yegnanarayana, "Analysis of mimicry speech", *Interspeech*, Portland, USA, September 2012.

[6] P. Gangamohan, V.K. Mittal, and B. Yegnanarayana, "A Flexible Analysis Synthesis Tool (FAST) for studying the characteristic features of emotion in speech", *Proc. 9th Annual IEEE Consumer Communications and Networking Conference - Special Session Affective Computing for Future Consumer Electronics*, Las Vegas, USA, pp. 266-270, 2012.

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

[8] 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, November 2008.

[9] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation", *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 14, pp. 972–980, May 2006.

[10] G. Fant, J. Liljencrants and Q. Lin,"A four-parameter model of glottal flow", *Quarterly Progress Status Report*, Speech Trans. Lab., KTH-Sweden, vol. 26, no. 4, pp. 001-013, 1985.

[11] P. Gangamohan, V.K. Mittal and B. Yegnanarayana,"Relative Importance of Different Components of Speech Contributing to Perception of Emotion", *Proc. Int. Conf. Speech Prosody*, pp. 657–660, May 2012.