# Detect Mimicry by Enhancing the Speaker Recognition System

**Soumen Kanrar and Prasenjit Kumar Mandal**

**Abstract** Mimicry voice sample is a potential challenge to the speaker verification system. The system performance is highly depended on the equal error rate. If the false accept to reduce, then the equal error rate decrease. The speaker verification process, verifies the claim voice is originally produced by the said speaker or not. The verification process is highly depended upon the biometric features carried out by the acoustic signal. The pitch count, phoneme recognition, cepstral coefficients are the major components to verify the claim voice signal. This paper shows a novel frame work to verify the mimicry voice signal through the two-stage testing. The first stage is GMM based speaker identification. The second stage of testing filters the identification through the various biometric feature's comparisons.

**Keywords** Mimicry voice · Speaker verification · Speaker identification · Acoustic signal · Phonetic · Biometric · Spectrogram pitch · Cepstrogram

## 1 Introduction

The sensitivity to computer by voice-altered impostor is using trainable speech synthesis technology makes the robustness of a speaker recognition system [1]. Mimicry voice is using synthetic speech against speaker verification based on the spectrum, pitch and cepstrum analysis. Pitch of a particular vocal cord of small duration extracted from the speaker's voice excitation. The same time duration is used to presents the movements of the acoustic signal corresponding to articulation, independent of language. Phonetic event changes significantly, and this is reflected the numbers of segments clearly visualize by spectrogram based on the different

S. Kanrar (✉) · P.K. Mandal
Vehere Interactive Pvt Ltd., Calcutta 700053, West Bengal, India
e-mail: Soumen.kanrar@veheretech.com

P.K. Mandal
e-mail: prasenjit.mandal@vehere.com

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 339,
DOI 10.1007/978-81-322-2250-7_3

range of frequency bands. Phonetic analysis is based on the important premise that it is possible to describe speech in terms of a sequence of segments. The crucial assumption, that each segment can be characterized by an articulatory target. 'Articulation' the activity of the vocal organs in making a speech sounds. The aforesaid biometrics offers greater security than traditional methods in person recognition by GMM based speaker identification [2–6]. In this paper, we present a robust approach to avoid the Mimicry voice that causes the potential security threat to the voice recognition system. We have seen that human tendency to copy the speaking style of some reputed personalities [1]. However, for the security point of view that mimicry voice used as the proxy of some existing voice model for any voice recognition system is a challenging issue. Mimicry voices are very vulnerable for any speaker recognition system [1]. The probable mimicry attack occurs in the domain of Voice dialing, banking over a telephone network, database access services, security control for confidential information and remote access of computers. So verification of the claim speaker is to identify the vocal track of the speaker from a number of existing speaker model present in the system [7]. Vocal cords are producing acoustic energy by vibrating as air passes between then. If the claim speaker voice is very nearer to an existing model in the system or numbers of models, then we proceed for second stage of verification. This is carried out by further speech analysis based on phonetic. Phonetic is concerned with the physical properties of acoustic signal. Phone is a unit of speech sound. Consonants and vowels are classified in terms of its place of articulation. Phonetic describes the place of articulation concentrates on a section through the mid line of vocal tract. Voice is the composition of sequence of discrete sounds or segments (Fig. 1). The segments are composed by consonants and vowels. The vowels and consonants are the fundamental part of the segmentation. The repeated opening and closing of vocal tract are syllables. More closed articulation is consonant, and more open articulation is a vowel. Consonant involve narrowing or restriction at an identifiable place in the vocal tract. The syllables often consider the phonological building blocks of words. All languages have different accent and other varieties of pronunciation, when sound is exemplified by a word in a particular language. If we choose the word for the second stage of verification, then the word should contain at least one vowel and the best chose of the word or words for both the speaker probably from the same language.

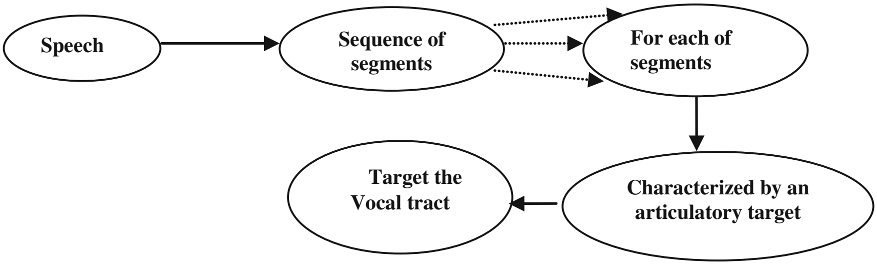

**Fig. 1** Speech segmentation

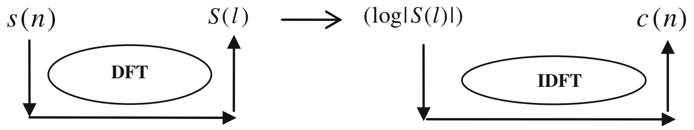

**Fig. 2** Cepstral coefficient processing

Neither the movements of the speech organ nor the acoustic signal offers a clear division of speech into successive phonetic units. The segmentation is influenced by knowledge of linguistically significant changes in sound.

Articulation is the mechanical or bio-mechanical process of vocal organ making the speech sound. Articulation is composed of excitation and vocal tract components. To analyze the articulation, these two parts have to be separated. The articulation is the convolution of the respective excitation sequence and vocal tract. Initially, the articulation is in the time domain. Initially, $s(n)$ is the articulation sequence in time domain, expressed as $s(n) = e(n) * u(n)$. Here $e(n)$ is the excitation sequence and $u(n)$ is the vocal tract sequence. In the frequency domain, it can be express as $S(l) = E(l) \cdot U(l)$.

The cepstral coefficient $c(n)$ is obtained through the number of process according to Fig. 2.

In the section of ‘target the vocal track’ via (Fig. 1) some types of biometric components used to match the individual’s features. Biometric is associated with a better degree of security and authentication. The stress in the sound and pitch count, excitation in the cepstrogram is the major biometric components used at the second stage of authentication. Acoustic display of a word in the second stage of verification is presented by the spectrogram. The pitch is one of the major constituent parts in the second stage of verification. Pitch presents how high or low, a voice sound seems. Obviously, the pitch count of the two different speakers of a particular word is considered for verification. The pitch depends (approximately) logarithmically on frequency of the acoustic signal. The cepstral analysis gives the excitation of the speaker in the acoustic signal for the particular speaking word by both the speakers. In the second stage of verification, we consider similar and very common words speaking by both speakers. Now by the spectrogram, pitch count and the cepstral analysis verify the claimed speaker voice with the nearest match vocal track present in the voice recognition system. In first stage of the speaker recognition system, the incoming speaker voices submitted to the system, and the second speaker voice is the known speaker voice, of which voice model is presented in the recognition system. Every speaker’s has a number of voice samples of that speaker forming a cluster of that known speaker. The total number of voice models present in the model list of the voice recognition system, which is sum of the number of clusters in the size of individual clusters. The first stage of comparison to the voice recognition system is purely automatic but the second stage of verification purely manually. The decision is based on both stages, first stage identifies the speaker, and second stage verifies the claim. In the first stage, the identification is done by the statistical hypothesis testing with the existing speaker model in the system. The testing is done by one to many testing where in the second stage of testing, it is one to one testing. The first stage of testing basically drills with the one to many matching, and the numerical score give the best probable prediction about the speaker with the list of the model present in the voice recognition system. The procedure for the first stage of checking is based on the Gaussian mixture model. The Gaussian mixture model is creating through the number of steps. The acoustic feature is extracted from the Mel Frequency Cepstral Coefficient (MFCC). Mel frequency Cepstral coefficients (MFCC) are collective build up the individual Mel frequency Cepstral (MFC). MFC is a physical representation of the short term power spectrum of an acoustic signal in a particular frequency band on a linear cosine transform of the log power spectrum [8]. The extracted acoustic feature from the voice signal after normalize produce various acoustic classes. These acoustic classes belong to an individual speaker voice or a set of speakers. The GMM is the soft representation of the various acoustic classes of an individual person voice or a set of speakers. The probability of a feature vector of being in the acoustic classes is represented by the mixture of different Gaussian probability distribution functions.

## 2 Model Development

Let us consider $X$ is a random vector i.e. $X = \{x_1, x_2, x_3, \ldots, x_L\}$ be a set of $L$ vectors, each $x_i$ is a $k$-dimensional feature vectors belong to the one particular acoustic class. $L$ is the number of acoustic classes and the vector $x_i$ are statistically independent. So the probability of the set $X$ for the $\lambda$ speaker model can be expressed as, $\log P(X|\lambda) = \sum_{i=1}^{L} \log P(x_i|\lambda)$. The distribution of vector $x_i$ with the $k$-dimensional components are unknown. It is approximately modeled by a mixture of Gaussian densities, which is a weighted sum of $l \leq k$ component's densities, which can be expressed as $P(x_s|\lambda) = \sum_{i=1}^{l} w_i N\left(x_s, \mu_i, \sum_i\right)$, $w_i$ is the mixture weight, where, $1 \leq i \leq l$ and $\sum_{i=1}^{l} w_i = 1$. Each $N\left(x_s, \mu_i, \sum_i\right)$ is a $k$ variate Gaussian component density presents as

$$N\left(x_s, \mu_i, \sum_i\right) = \frac{\mathrm{e}^{-\{0.5(x_s-\mu_i)' \sum_i^{-1} (x_s-\mu_i)\}}}{(2\Pi)^{k/2} |\sqrt{\sum_i}|},$$

$\mu_i$ is the mean vector and $\sum_i$ is the covariance matrix. $(x_s - \mu_i)'$ is the transpose of $(x_s - \mu_i)$.

In the speaker identification from the set of speakers $\{S_i\}$ where $i$ is countable finite and $X$ is given utterances, if we claim that the utterance produce by the speaker $S_k$ from the set of speakers $\{S_i\}$. So the basic goal is how it is a valid claim that the speaker $S_k$ makes the utterance $X$. The utterance $X$ is a random variate that follows the Gaussian mixture probability distribution. The claim follows the expression $P(S_k/X)$ present the probability of the utterances $X$ produce by the speaker $S_k$. So $P(\bar{S}_k/X)$ is the probability that the utterances, $X$ is not produced by the speaker $S_k$. Let, $\bar{S}_k = \bigcup_i S_i - S_k$ is the collection of large heterogeneous speakers from different linguistics, including both genders and from different zones of the globe. $\bar{S}_k$ can be better approximated as universal model or world model. It is presented as $\bar{S}_k \approx \omega$ (say). Now the claim be true according to the rule,

$$\text{if } P(S_k/X) > P(\bar{S}_k/X) \text{ then the utterance produce by } S_k \tag{1}$$

else, the claim is false. So, the utterance produce by other speaker, except $S_k$. Since, it is a probabilistic prediction about the claim. However, the process can't predicate the certain events, with values 0 or 1. According to the general definition of probability, produce highest level of prediction about the claimed speaker with the numeric values. It is very often that this predicated score very much depends upon the acoustic classes that obtained from the long step procedure. So the extracted feature largely depends on the digitalization of analogue acoustic signal. There are high chances that the claimed speaker voices, probability comparison values may not be the best or highest value lie in the interval (0, 1).

By the Bayes theorem the expression (1) produce

$$\frac{P(X/S_k)P(S_k)}{P(X)} > \frac{P(X/\omega)P(\omega)}{P(X)},$$

since we assume that $X$ is not silence clearly, $P(X) \neq 0$.

We get,

$$\frac{P(X/S_k)}{P(X/\omega)} > \frac{P(\omega)}{P(S_k)} = \lambda_k \approx \lambda \tag{2}$$

$\lambda$ is a pre assume threshold. To compact the all possible predication we consider the log on the both sides [9].

$$\log\frac{P(X/S_k)}{P(X/\omega)} > \log\lambda_k = \lambda'_k \tag{3}$$

The predicated values indicated how closer the claimed speaker to the existing speaker's voice after comparison. The predicated values are Gaussian in nature so further compactness be done on the predicated values by the statics [10, 11].

$$\frac{\frac{P(X/S_k)}{P(X/\omega)} - \mu}{\sigma} > \lambda \tag{4}$$

## 3 Simulation Results and Discussion

The comparison testing being done in two stages, at the first stage, we consider two speakers voice one of the voice from Indian film actor and personalities Amitabh Bachchan and other voice of the comedian Raju Srivastav. Initial stage, voice models of both the speakers present in the voice recognition system along with the other voice models. In the voice model list, there are 50 voice models present with their model identifier number. The model identifier number of pure Amitabh Bachchan Voice is 1, Model identifier number for Comedian Raju Srivastav pure voice is 2. Model identifier number 3, 4, 5, 6, …, 50 are the different Bollywood male film star voice models. We have taken two new input voices one of Mr. Amitabh Bachchan and other voice of Comedian Raju Srivastav. Figure 3 shown the predicated test score presented by solid line and dash line for speaker Amitabh Bachchan and Comedian Raju Srivastav. The predicated score for Amitabh Bachchan voice matches with Amitabh Bachchan voice with score value 1.6 and with Raju Srivastav Model with predicated score value 1.65. The given voice of Amitabh Bachchan match with the model identifier 3 and 4 with a score values 0.8 and 0.6. If we consider the accept level of prediction value 1.5, then Mr. Bachchan voice match with both the speakers Mr. Bachchan Model and Raju Srivastav model. The line dash line indicated the Comedian Raju Srivastav new voice. The predicated score of the new voice of Raju Srivastav matched with Mr. Bachchan voice model with score value 1.5 and with Raju Srivastav voice model with predicated score value 2.3. The predicated match score with another model identifier 3 is 0.4. If we consider the accepted range of the prediction is [1.5, +∞), then Amitabh Bachchan voice matches with Amitabh Bachchan voice as well as Raju Srivastav voice.

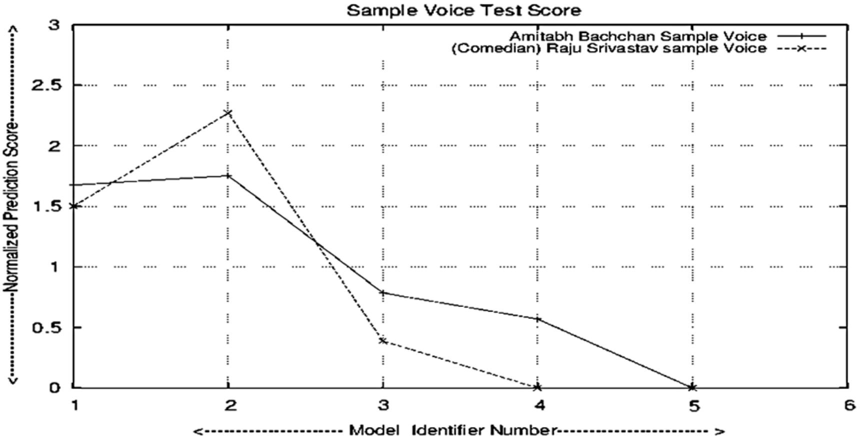

**Fig. 3** Simple voice score

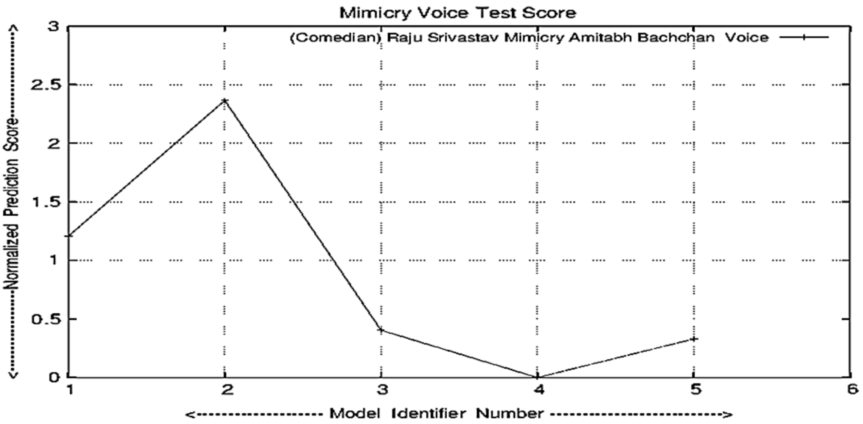

**Fig. 4** Mimicry voice score

Figure 4 shows that Raju Srivastav new voice only matches with Raju Srivastav existing model. Other matches with the existing model identifier 3 and model identifier 5 are 0.45 and 0.4 respectively. If we consider the level of acceptance is 1.5 clearly, the Raju Srivastav Mimicry voice matches only with Raju Srivastav existing voice model.

The problem arises, if the level of acceptance is considered as 1.0, then mimicry voice matches with Amitabh Bachchan as well as Raju Srivastav. The above comparative predicated score indicated that mimicry voice is an original voice of Raju Srivastav, not the voice of Amitabh Bachchan. But there is little chance that the voice may be Amitabh Bachchan voice according to the level of acceptance. The Second level of verification results is shown in Figs. 5, 6, 7, and 8. In Figs. 5 and 6 present the wave form, spectrogram and the pitch counts of the two incoming voice in the very compact form. The first row presents the Pitch count, second row presents the spectrogram and the third row presents the acoustic signal of the speaker. We select the word spoken by both the speaker which contains at least one vowel and in the same language. Here we manually selected segment portion of the spoken-word ‘Sign’ of both the speaker. Figures 5 and 6 present the comparison of both the speaker’s pronunciation of the word ‘Sign’. Figure 5 present pronounce of the word ‘Sign’ by Amitabh Bachchan and the Fig. 6 presents pronounce of the word ‘Sign’ by comedian Raju Srivastav. The number of pitch counts for the speaker change, 19 for Amitabh Bachchan spoken word and 15 for Raju Srivastav spoken word. The spectrogram of Amitabh Bachchan spoken word largely changes with the Raju Srivastav spoken word. At the frequency band 3,000–3,500, the harmonics are clearly changes.

Figures 7 and 8 present the Cepstral of the 3,000 ms short term speech segment of the spoken-word ‘sign’. X axis is the time axis, and the Y axis is the cepstral axis. Since cepstral is derived from the log magnitude of the liner spectrum, so it is also symmetrical in the cepstral domain.

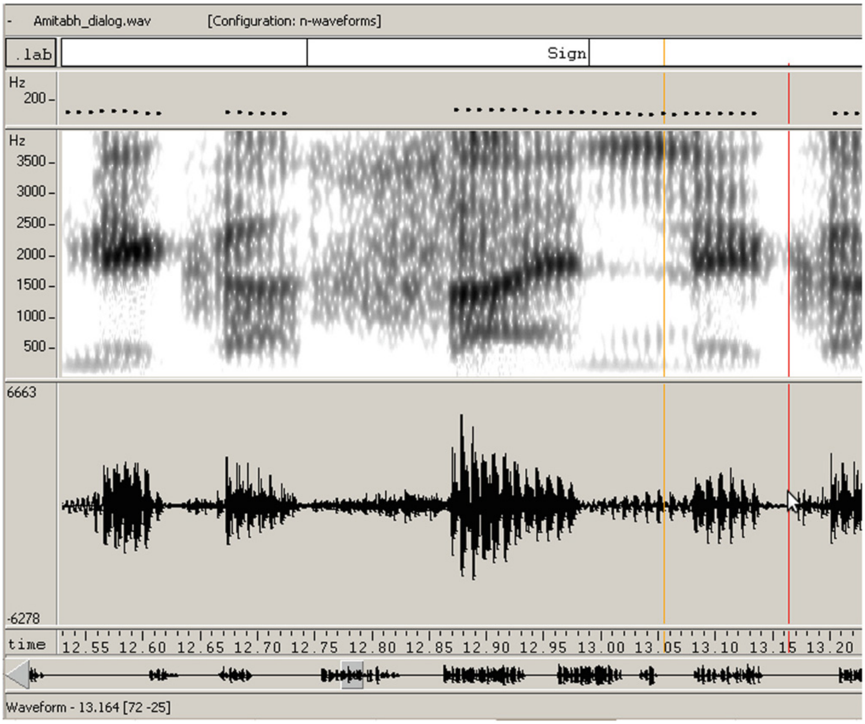

**Fig. 5** Spectrogram of word

Figures 7 and 8 present one symmetric part of cestrum. In the cepstral domain, the vocal tract components are represented by the slowly varying components concentrated near the lower cepstral value area i.e. along the Y-axis. Where the excitation of the speaker's voice is fast varying concentrated near the higher cepstral value area along the Y-axis.

The comparative study of the Figs. 7 and 8 has shown clear difference between them. Figure 7 is the cepstral presentation of the word 'sign' spoken by Amitabh Bachchan and Fig. 8 present cepstral for the word 'sign' spoken by Comedian Raju Srivastav. During the time interval [0, 1,000] the cepstral asset value for Amitabh Bachchan voice is much higher than the Comedian Raju Srivastav voice. Futhermore, we have notice that excitation presents in the interval [1,200, 1,400], [1,550, 1,800], [2,200, 2,500], [2,600, 2,900] for the Amitabh Bachchan voice. In Comedian Raju Srivastav spoken word 'Sign' this many amounts of excitation are not present during the interval [1,200, 3,000]. Further, we have notice that during the interval [750, 3,000] vocal tract present in Comedian Raju Srivastav spoken word 'sign' is much higher than the spoken-word 'sign' by Amitabh Bachchan. Based on the above comparison in the second stage of verification, firmly we can come to the

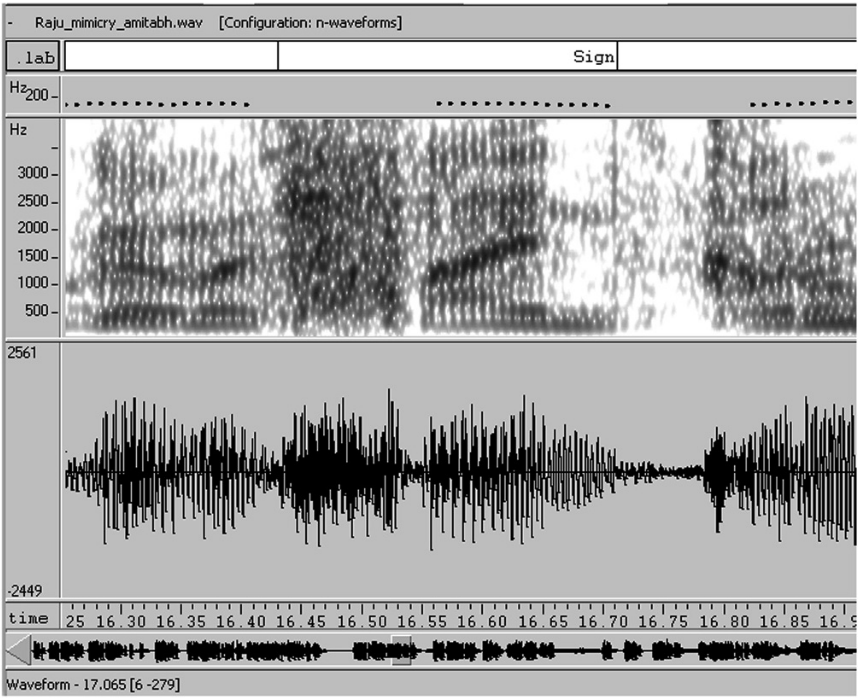

**Fig. 6** Spectrogram of word

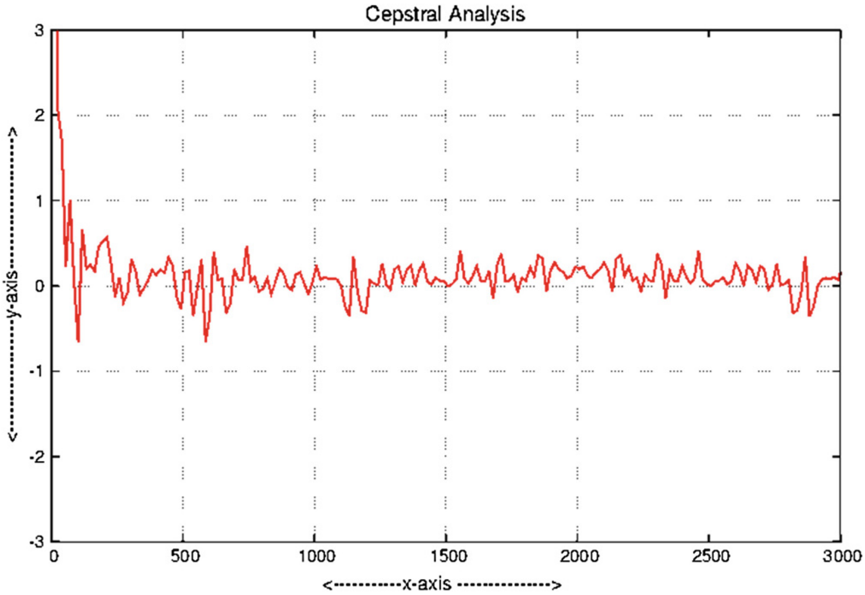

**Fig. 7** Cepstrogram of word

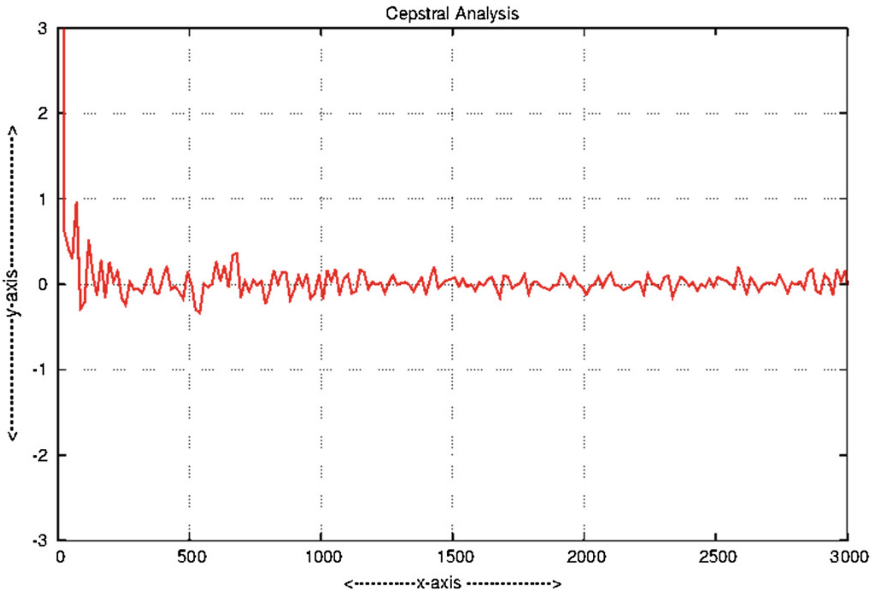

**Fig. 8** Cepstrogram of word

conclusion. The articulation comes from different speakers, although the mimicry voice of comedian Raju Srivastav has spoken the same context with same speaking style of Amitabh Bachchan.

## 4 Conclusion

This work presents the work flow of two stages verification for the mimicry voice testing. At the first stage, GMM based speaker identification, which is a one to many identification processes. The second stage is the phonetically based speaker verification for very closer identified speaker. This work to be tested for a large number of collected mimicry voice samples in further extension of this work.

## References

1. Lee L.W., et al.: Vulnerability of speaker verification to voice mimicking. In: Proceeding of International Symposium on Intelligent Multimedia, Video and Speech Processing. Hong Kong, Oct 2004, pp. 145–148
2. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digit. Sig. Process. **10**, 19–41 (2000)

3. Xiang, B., et al.: Short-time gaussianization for robust speaker verification. In: Proceedings of ICASSP, 2002, vol. 1, pp. 681–684 (2002)
4. Reynolds, D.: Automatic speaker recognition using Gaussian mixture speaker models. Linciln Lab. J. **8**(2), 173–191 (1995)
5. Reynolds, D.: Speaker identification and verification using Gaussian mixture speaker models. Speech Comm. **17**, 91–108 (1995)
6. Reynolds, D.: Channel robust speaker verification via feature mapping. In: Proceedings of International Conference on Acoustics Speech Signal Process, pp. 53–56. (2003)
7. Apsingeker, V.R., DeLeon, P.: Speaker model clustering for efficient speaker identification in large population applications. IEEE Trans. Audio Speech Lang. Process. **17**(4), 848–853 (2009)
8. Bimbot, F., et al.: A tutorial on text-independent speaker verification. EURASIP J. Appl. Sig. Process. **4**, 430–451 (2004)
9. Morrison, G.S.: Measuring the validity and reliability of forensic likelihood—ratio system. Sci. Justice **5**, 91–98 (2011). doi:10.1016/j.scijus.2011.03.002
10. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalizing for text-independent speaker verification system. Digit. Sig. process. **10**, 42–52 (2000)
11. Mirghafori, N., Heck, L.: An adaptive speaker verification system with speaker dependent a priori decision thresholds. In: Proceedings of ICSLP, Denver Colorado, Sept 2002