International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-4 Issue-13 December-2013

Automatic Identification of Modal, Breathy and Creaky Voices

Poonam Sharma[1], Ajay Sharma[2]

Abstract

This paper presents a way for the automatic identification of different voice qualities present in a speech signal which is very beneficiary for detecting any kind of speech by an efficient speech recognition system. Proposed technique is based on three important characteristics of speech signal namely Zero Crossing Rate, Short Time Energy and Fundamental Frequency. The performance of the proposed algorithm is evaluated using the data collected from three different speakers and an overall accuracy of 87.2 % is achieved.

Keywords

Zero Crossing Rate, Short Time Energy, Fundamental Frequency, Cepstrum

1. Introduction

After decades of research and work many Automatic Speech Recognition Systems have been developed but almost all of them work under constrained environment and when exposed to other environment their recognition performance degrades. One of the main reasons for the performance degradation is voice quality. Almost none of them take into consideration the voice quality of the speaker. The term voice quality describes the quality of sound produced with a particular setting of the vocal folds which includes modal phonation that is produced by a plain or normal voice quality and the non- modal phonation of breathy and creaky voice qualities. If we can reliably extract acoustic features that differentiate phones that differ from each other regarding voice quality, then such a difference can be modelled in a Speech Recognition System with the goal of improving recognition performance. In recent years considerable efforts has been spent by researchers in solving the problem of classifying speech Signal according to the quality of voice. Various methods based on features of speech like turbulent noise, waveform peak factor, Harmonics-to-Noise Ratio (HNR) etc. [1 and 2].Classification and Regression Tree (CART) analysis and various high order statistical features were also used for this classification [3]. But still for the efficient performance of voice quality dependent recognizers, more effective methods are required. Malyska and Quatieri (2008) introduced a general signal-processing framework for interpreting the effects of both stochastic and deterministic aspects of non-modality on the short term spectrum. They showed that the spectrum is sensitive to even small perturbations in the timing and amplitudes of glottal pulses. In addition, they illustrated important characteristics that can arise in the spectrum, including apparent shifting of the harmonics and the appearance of multiple pitches. Lee et al. (2008) proposed Higher-Order Statistics (HOS) based features to improve classification performance of voice quality measurement. These features were means and variance of kurtosis and skewness which showed meaningful differences in normal, breathy and rough voices. They also used conventional features like jitter, harmonic to noise ratio. They used Classification and Regression Tree (CART) analysis and by utilizing both conventional and HOS based features they were able to get an 89.7% classification rate [8]. The first voice quality dependent speech recognizer was developed by Yoon et al. (2008). They used both spectral and temporal features specifically the harmonic structure and the mean autocorrelation ratio. These features were used by SVM (Support Vector Machine) to classify different type of voice qualities and it was found that Voice quality distinction reflected in PLP coefficients. They obtained 69.23% classification accuracy where the baseline accuracy was 50% a 19% improvement was found which in turn suggests that they could conduct a speech recognition experiment that utilizes the voice quality information, using PLP coefficients as input feature vectors. Finally they used HMM for recognition where first they developed a baseline system and tested its accuracy which acted as the baseline accuracy and then they incorporated the voice quality features into the HMM and tested it. Finally they found that in there was slight increase in the accuracy of the detection [7].

Poonam Sharma, CSE Department, Sharda University, Greater Noida, India.
Ajay Sharma, Samsung Research Institute, Noida, India.

The method we used in this work is a simple and fast approach and can overcome the problem of classifying the speech. In section 2 we discuss various features and the facts observed from them. In section 3 the proposed work and algorithm is described and in section 4 results are being discussed.

2. Signal Features and Observations

A. Zero Crossing Rate

The Zero Crossing Rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. For this work ZCR was calculated using a window of 20 ms and it was observed that for both modal and creaky voice qualities the ZCR was less than 0.1 but for breathy voice and noise it was above 0.1 and less than 0.3 and for voiceless sounds it was above 0.3 as shown in Fig. 1,2 and 3 for different voice qualities.

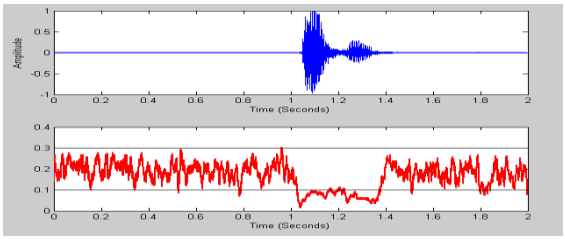

Fig. 1. Plot of "bahr" spoken in modal voice quality with its ZCR .

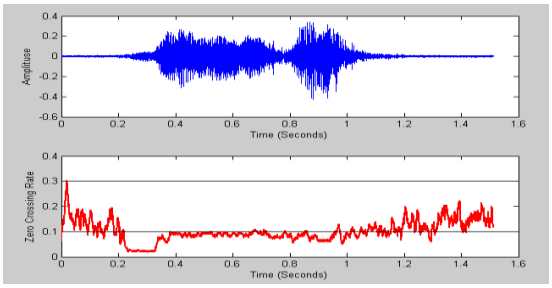

Fig. 2. Plot of "bahr" spoken in Creaky voice

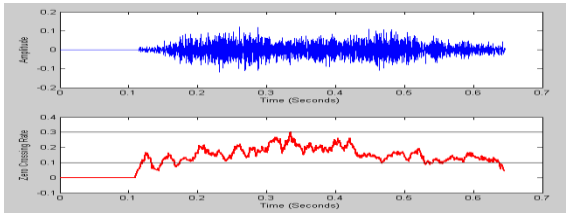

Fig. 3. P lot of "bahr" spoken in breathy voice quality with its ZCR

B. Fundamental Frequency

Fundamental Frequency (F0) or pitch is defined as the frequency at which the vocal cords vibrate during a voiced sound. Basically, there are two categories of approaches for pitch tracking. One category is in the time domain, and the other category is in the frequency domain. Time-domain analysis could use some time-related features such as ZCR (Zero-Crossing Rate), peak picking, and autocorrelation. Frequency domain analysis could apply, for example, to cepstrum and harmonic matching [4]. For this work two approaches were tested Time domain approach using autocorrelation and Spectral domain approach using cepstrum with hamming window of 40ms as shown in Fig. 4 and 5 and it was observed that cepstrum based approach was more effective as it was only taking into consideration the voiced region of the audio signal being produced.

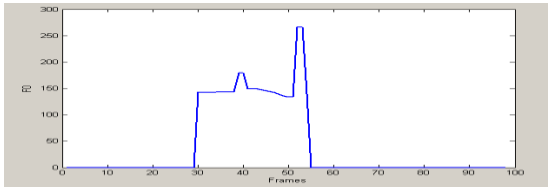

Fig. 4. F0 of "shalgam" using autocorrelation

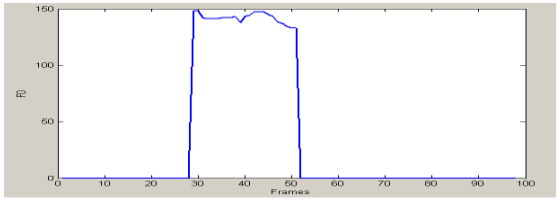

Fig. 5. F0 of "shalgam" using cepstrum

To differentiate between the modal voice and creaky sound fundamental frequency estimation (also called pitch detection) is used. It is well known that F0 of creaky sound is always less than modal sound because for modal voice vocal fold length may be considered to be medium with the length increasing as fundamental frequency (F0) increases and for vocal fry (creaky voice) vocal fold length is short which causes the F0 to decrease from that of modal voice [1].

After taking the average values of F0 for various words it was observed that that the F0 of creaky sound always lied in the range of 95 to 135 Hz and for modal voice it was above 135 Hz.

C. Short Time Energy

Breathy sound and noise have almost same zero crossing rate so using zero crossing rate they are inseparable. But the STE of noise is very less than of voiced region and breathy sound lies in the voiced region [5 and 6]. This property of breathy sound is used for the identification of breathy sound.

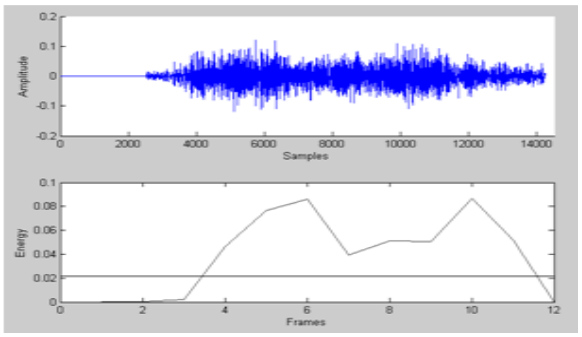

Fig. 6. Short Time Energy of "bahar" spoken in Breathy voice

3. Method for Identification

After analyzing the results from different features calculated the algorithm was designed for identifying the voice quality.

A. Algorithm

Step 1: Read the sound file and create speech vector $X(n)$ where n is the length of speech signal.
Step 2: Take 20 ms rectangular window and calculate ZCR vector (ZC) and map it to the same length as that of speech signal.
Step 3: Compute Fundamental Frequency Vector (F) taking hamming window of size 40 ms.
Step 4: Map F0 vector to the vector of length similar to that of speech signal.
Step 5: Compute Short Time Energy vector (STE) taking Hamming window of 50 ms.
Step 6: Map STE vector to the vector of length similar to that of speech signal.
Step 7: Calculate threshold T_E for STE.
Step 8: Make output vector (OUT) of length equal to X and initialize all its values to zero.
Step 9: Repeat for i=1 $to\ n$
If $ZC(i)$ < 0.1
If $F\ (i)$ > 145 then
set $OUT(i)$=0.1 (for modal sound)
else
set $OUT(i)$=0.3 (for creaky sound)
end if
Else if $ZC(i)$ between 0.1 and 0.3 then
If $STE(i)$ >T_E then
then set $OUT(i)$=0.2 (for breathy sound)
Else if $ZC(i)$ > 0.3 then
set $OUT(i)$=0.1 (for modal sound)
End if
Step 10: plot X and OUT.

B. Flowchart

Fig. 7 shows the flowchart of the algorithm designed. The input for the algo is a speech vector X(n) and based upon the features the algo decides that the quality of voice.

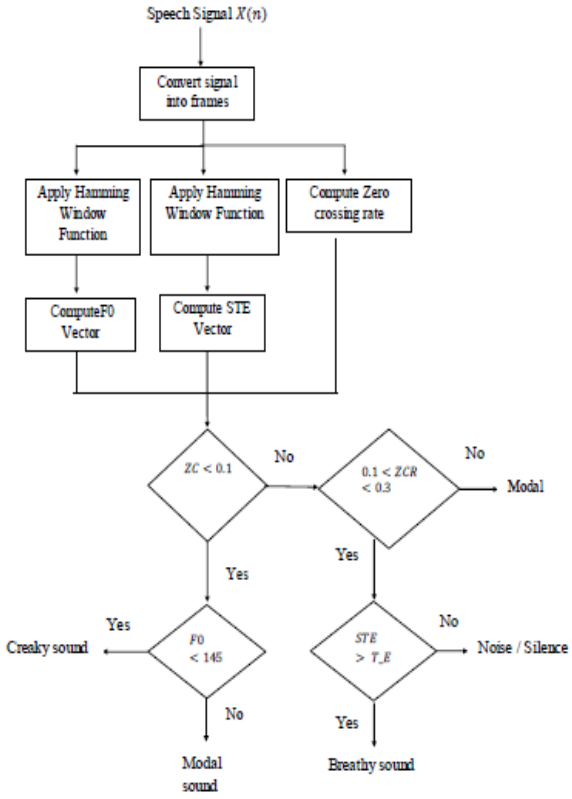

Fig. 7. Flowchart of the Algorithm designed

4. Result

After applying the algorithm discussed the output was in the form of a matrix and was of the same length as that of the length of the speech signal. The simplest output of the algorithm is shown in Fig. 8 and 9. The blue line is the output. The part where this line lies on value 0.3 means it is an silence zone, similarly 0.2 values means breathy voice, 0.4 creaky voice and finally 0.1 means modal voice.

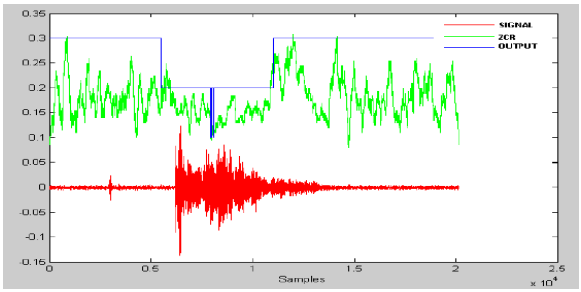

Fig. 8. Output of algorithm for "ghar" spoken in breathy voice.

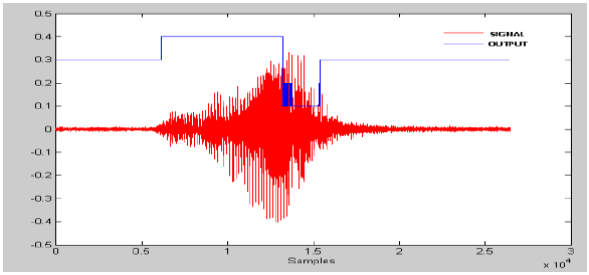

Fig. 9. Output of algorithm for "ghar" spoken in creaky voice.

The accuracy is calculated as

$Accuracy\ (in\ percentage) = (No.of\ samples\ correctly\ detected*100)/Total\ voiced\ samples$

Table I. shows the average accuracy obtained for 10 words spoken 3 times each in modal voice.

As it is clear from the above table high degree of accuracy is achieved in identifying modal voices.

TABLE I. Accuracy (in percentage) obtained from algorithm for modal voice.

Sr. No	Word	Accuracy(in percentage)
1	"ajay"	91.17
2	"ghar"	89.03
3	"bahr"	92.16
4	"kabutar"	87.73
5	"shalgum"	85.56
6	"aag"	95.03
7	"chabi"	89.56
8	"chori"	86.67
9	"gamla"	95.6
10	"ghas"	94.35
Overall accuracy (in percentage)		91

Almost 91% of the samples are correctly identified as modal sounds. Accuracy up to 95% is achieved for some words and for some words it is a bit low, i.e., 85%.Table II. Shows the average accuracy obtained for 10 words spoken 3 times each in creaky voice.

TABLE II. Accuracy (in percentage) obtained from algorithm for creaky voice.

Sr. No	Word	Accuracy (in percentage)
1	"ajay"	75.6
2	"ghar"	75.73
3	"bahr"	80.4
4	"kabutar"	69
5	"shalgum"	77.9
6	"aag"	84.93
7	"chabi"	79.37
8	"chori"	91.23
9	"gamla"	79.4
10	"ghas"	83.6
Overall accuracy (in percentage)		80

The accuracy obtained for creaky voice is less than that for modal voice. This is due to the mixed nature of creaky voice ,i.e., even if the speaker speaks in creaky voice some phonemes like /ey/, /r/, /ta/ etc. are not creaky. For words like "kabutar" where the phoneme /ta/ is extended for a long time the accuracy rate is 67.5%. The average accuracy obtained is 80.72%. The accuracy can be improved by considering average F0 of whole word (not the frames) but then the algorithm will work for audio file containing only one word. So there is a trade-off between accuracy and multiple words in a file.

Table III. Shows the average accuracy obtained for 10 words spoken 3 times each in Breathy voice.

TABLE III. Accuracy (in percentage) obtained from algorithm for breathy voice.

Sr. No	Word	Accuracy (in percentage)
1	"ajay"	89.5
2	"ghar"	91.62
3	"bahr"	88.52
4	"kabutar"	89.26
5	"shalgum"	91.35
6	"aag"	95.68
7	"chabi"	88.84
8	"chori"	90.09

9	"gamla"	96.26
10	"ghas"	94.56
Overall accuracy (in percentage)		89

For breathy sound the overall accuracy obtained is 89.84% , but the error obtained in case of breathy is evenly distributed over the whole range of samples. Also for some words the nature of phoneme changes to unvoiced for e.g., in case of "kalam" the /m/ phoneme becomes unvoiced which causes the accuracy rate for this word to decrease. The overall accuracy of the algorithm for the collected data from three different speakers is shown in Table 1V.

TABLE IV. Accuracy of proposed Algorithm

Sr. No	**Voice Type**	**Accuracy (in percentage)**
1	Modal	91.05
2	Creaky	80.72
3	Breathy	89.84
Overall accuracy (in percentage)		87.2

5. Conclusion

After decades of work in the domain of speech recognition, voice quality still remains one the untouched part where very less work has been done and very less accuracy has been achieved. This thesis is an effort to improve the accuracy to identify different voice qualities. The algorithm proposed in this work has successfully been able to identify the different voice qualities with some error rate especially in case of creaky voices. The algorithm was tested for 2 male speakers speaking in three different voice qualities (modal, breathy and creaky). For modal voice accuracy of 91.05% was achieved, for creaky it was a bit less ,i.e., 80.72% and finally for breathy voice it was 89.84%. The overall accuracy of proposed algorithm is 87.2% which is quiet good considering the previous works. The results achieved in present study motivate to extend the present work to achieve higher degree of accuracy especially in case of creaky voices where the algorithm shows less accuracy and also for breathy and modal voices more work can be done to achieve more accuracy in harsh environment and make the identification more robust.

References

[1] Childers, D. G. and Lee, C.K., 1991. Vocal quality factors: analysis, synthesis, and perception. Journal of the Acoustical Society of America . vol. 90, no. 5, pp. 2394-2410.

[2] Yoon, T., Zhuang X. , Cole J. Johnson M., 2009. Voice quality dependent speech recognition. Linguistic Patterns in Spontaneous Speech (Language and Linguistics).

[3] de Krom, Guus. "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals." Journal of Speech, Language, and Hearing Research 36, no. 2 (1993): 254-266.

[4] Lee, J. Y., Jeong, S., Hahn, M., Choi, H. S., 2008. Automatic voice quality measurement based on efficient combination of multiple features. International conference on bioinformatics and biomedical engineering, vol. 4, pp. 2583-2586.

[5] Zhao, X., O'S haughnessy, D. and Minh-Quang, N., 2007. A processing method for pitch smoothing based on autocorrelation and cepstral F0 detection approaches. International.

[6] Atal, B., Rabiner, L., 1976. A Pattern Recognition Approach to Voiced-Unvoiced-S ilence Classification with applications to Speech Recognition.IEEE Transaction on Acoustics, Speech and S ignal P rocessing, vol. 2, no. 2, pp. 201-212.

[7] Gordon, M., 2001. Linguistic aspects of voice quality with special reference to Athabaskan. Proceedings of the 2001 Athabaskan Languages Conferenceograph Series).

[8] Malyska, N. and Quatieri, T. F., 2008. Spectral representations of nonmodal phonations. IEEE transactions on audio, speech and language processing, vol. 16, no. 1, pp. 34-46.

Ajay Sharma, Born Kangra (Himachal pradesh) 20th May 1988. B.Tech (IT) from Kurukshetra Univeristy, M.tech (CSA) Thapar University. Currently working as Engineer in Samsung Reasearch Institue, Noida.

Poonam Sharma, Born Kurusketra (Haryana) 22nd March 1989. B.Tech (CS) from Kurukshetra Univeristy, M.tech (CSA) Thapar University. Currently working as Assitant professor in Sharda University, Greater Noida.