Recognition of An Articulation From The Speechless Video

Lavanya K J[1], Praveen K B[2], Dr. Sivakumar B[3]
[1]Dept of Telecommunication Engineering
[2]Assistant Professor, Dept of Telecommunication Engineering
[3]Professor, Dept of Telecommunication Engineering
[1, 2, 3] Dr.Ambedkar Institute of Technology, Bengaluru

Abstract- *The paper proposed is dealing with producing a speech in the form of text from a person talking visual. K means algorithm is used for generating audio speech.*

First two techniques are proposed, Regression method and classification method, were proposed for estimating the static envelope functions with the lively appearance of the model with its visual features. Speech excitation data were now not available in the visual signal, and then feature and model level techniques are brought to artificially generate fundamental frequency and aperiodicity. Then those parameters are mixed with the straight vocoder to generate speech, using MATLAB software program.

Prototype model including MFCC (Mel Frequency Cepstral Coefficients) and Vector Quantization is carried out, all the test files are stored in the database for recognition purpose means depending on how speaker talks, speed and accent..etc. Signals are multiplied with Mel space filter bank to get Mel frequency wrapping, which will increase the resolution and MFCC is being processed, with the help of acoustic vector, the test signals are compared with trained signals to recognize the signal. Use of Kmeans algorithm will increase the accuracy as it is containing outer fittings and inner fittings in classification handled in Kmeans.

Keywords- Speech processing, Feature extraction, Feature collection, MATLAB

I. INTRODUCTION

The project work is related to generating an intelligible audio speech in the form of text from video of person's talking.

The system is applicable in domain of silent speech interfaces which transform information from non-acoustic sensors into an audio signal with electromagnetic articulography (EMA), ultrasound and electromyography (EMG).Silent speech interfaces are used in the areas like providing an artificial voice for patients who have undergone a laryngectomy. Other area like surveillance, where a video signal of a speaker is available but audio is not present, because of the distance to the target or may be audio recording device is not present. Giving privacy to the cellular telephone users with silent speech input is one of its applications. For providing real-time conversations support, silent speech interfaces should be operated fast such a way that delay should not be more than 150 ms.

This approach will work to exploit the correlation between visual and audio speech. The proposed approach to reconstructing audio speech from visual speech is based on obtaining the necessary parameters to drive speech production model. This model is requiring acoustic features comprising vocal tract (spectral envelope) and excitation parameters, like fundamental frequency, a measure of aperiodicity and a non-speech/unvoiced/voiced decision.

In conventional speech coding applications, these features are extracted from the audio signal. From correlation analysis, it is apparent that certain acoustic features will not be available from the visual stream, such as excitation, while some level of spectral envelope information is present.

II. METHODOLOGY

1) Speech recordings

To train and test the speech recognizer, it should be recorded from scratch a total of 41 speech sentences that comprised of 18 predefined names and were recorded on a high quality microphone and should be processed in Adobe Premiere Pro to minimize any noise occur for a clean control test, and parallel noise can be easily introduced to know its effects on recognition performance. The audio was stored as 16 bit/44.1kHz wave files.

2) Training recordings

20 training recordings were made by speaking out the 18 names in identical order 20 times, with the subject speaking naturally with slight variations of the voice. This is the speech,

which has to be replicated for the highest word accuracy during the demonstration.

3) Testing recordings

An additional 21 testing recordings were made in a similar fashion with variations of the number of names, order of names, articulation speed and tone. In this way this can provide guarantee that speech recognizer works with certain amount of flexibility.

4) Live recordings

The speeches are available for the live demonstrations were recorded on the same microphone using Adobe Audition with no post processing.

One thing to know is that recognition system was unable to pick out any vocabulary when it was recorded on a different microphone with similar high quality playback (more will be explained in the Results Evaluation part of the report).

5) Task grammar and word network

For HTK (Hidden Markov Model Toolkit), HTK is Toolkit which is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition research although it is used for numerous other applications including research into speech synthesis, character recognition and DNA sequencing.

HTK is in use at hundreds of sites worldwide. To understand how the names in each sentence are spoken, a wordnetwork needed to be defined using HTK's Standard Lattice Format (SLF) in which each word to word transition is stated.

To make things easier, a grammar file is made with the following notation ([silent] < $name > [silent])that represents a random alternate name between optional silences at both start and end of the word. Finally, it produces the word network by running the grammar file through HPARSE.

6) The Dictionary

The dictionary defines how names in each sentence are pronounced as a sequence of phones, but since the dataset is very manageable, it just specifies the training model name for each word entry:

a) Silent Silent
b) Michael
c) Michael Jack
d) Jack
e) Ben Ben
f) Change
g) Out

7) Label files

To train the HMM for each vocabulary, Need to create label files for each recording that associated with each name entity with the waveform at specified time segment in the recording. Praat (Praat is an open-source program for the analysis of speech in phonetics, created by Paul Boersma and David Weenink of the University of Amsterdam), was used to annotate the name entities with ease and then a C# program written by us is used to batch convert Praat's Text Grid format to HTK's Master Label File (MLF) format.

III. FEATURE EXTRACTION

The implementation of Mel Frequency Cepstral Coefficients (MFCC) feature extraction is achieved through calling a custom MATLAB function feature Extraction (sample, fs, filename, channelsNum) and providing four parameters:

- sample - The speech signal vector to beprocessed.
- fs - Sampling frequency of the speechsignal.
- filename - Name of MFCC matrix with feature vectors as columns tooutput.
- channelsNum - Number of filterbankchannels.

1) Resampling the audio file

The sample frequency of the original recordings was 44.1 kHz but the test noise file were given were sampled at 16 kHz therefore resampling was necessary.

Apart from the fact that resampling was required in order to add noise to the original recordings, also discovered that the results were obtained with the speech recognizer are improved when the clean audio files are down sampled from 44.1 kHz to 16 kHz.

2) Splitting the audio file into frames

The feature extraction uses 20ms frames with a 10ms overlap. The function loops through the whole audio file and extracts a frame every overlap period.

There is a possibility of further testing and evaluation with different frame sizes that could lead to potentially

interesting results but due to time restrictions we were not able to conduct these tests.

3) Pre-emphasis

Initially, set out to design a simple high-pass filter by following the example set in the lecture slides. Our implementation, however, did not result in an expected increase in the accuracy of the speech recognizer

Therefore, it uses the built-in MATLAB filter () function filter ([1,-0.97], 1, frame) with alpha of 0.97 as specified in the lecture slides.

4) Hamming window

Using the built-in hamming function and create a hamming window which in turn is applied to the current frame.

The use of a hamming window smooths out the amplitude of the discontinuities at the boundaries of each frame reducing the spectral leakage that occurs when applying a Fourier Transform on a finite record of the signal (in our case a 20ms frame).

5) Spectral analysis

For speech recognition and feature extraction we are interested in the frequency domain of the signal. To obtain that data we apply a Fast Fourier Transform (using the built-in fft() function) to the current frame.

The resulting complex spectrum is then converted into a frequency magnitude spectrum (using the built-in mag() function). Since the values of the magnitude spectrum are mirrored we only use the first half of the array.

6) Linear filter bank

The next step in feature extraction is obtaining the magnitude coefficients. Ideally, our feature extraction system would use the melscale filter bank which imitates the frequency response of a human ear but due to time limitations we obtain the value of energy in each band through a rectangular, linearly-spaced filter bank.

The function takes number of channels as a parameter and by testing different values we found that our speech recognizer produces the best results when the linear filter bank uses 100 channels to generate the coefficients (more in the Results Evaluation part of the report). The dynamic range of the coefficients is then reduced by applying a log function to the result.

7) Estimating the frequency

In order to isolate the vocal tract information from the excitation information the feature extraction system converts the coefficients from the spectral representation to the cepstral representation. This is achieved through a Discrete Cosine Transform (using the built-in dct() function). The resulting quefrency estimate is truncated in half to only retain the useful vocal tract data.

8) Result

The result of looping through the audio file is a 2-D matrix containing a feature vector for each analyzed frame.

The number of feature vectors corresponds to half of the number of channels used in the linear filter bank. The matrix is then written to a file format recognized by HTK.

IV. RESULTS AND DISCUSSION

There are three stages for obtaining an output in recognition of audio speech from a video of person talking, they are,

1) Feature collection
2) Select query
3) Classify

1) Feature collection

In this first stage, the system will be trained before the execution, many audio signals are trained and stored in the data base, and is shown below figure

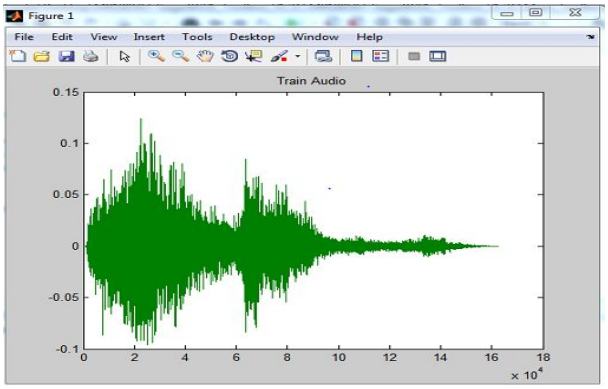

Fig: feature collection for database

At this stage, all the audio files are uploaded in to the database and these audio files are compared with the video of person talking to generate audio files.

2) Select query

At this stage, the audio files are already trained and stored in the database, and the video files of person talking are tested with the already stored audio files, and this is shown in the below figure

Fig: selecting video file

At this stage, the video file which was selected for testing will be compared with the already stored audio files of database and the audio file which matches with the selected testing video to produce an output.

3)Classify:

This is the last stage, where the audio files are extracted from the video of person talking; this will be shown in the figure below

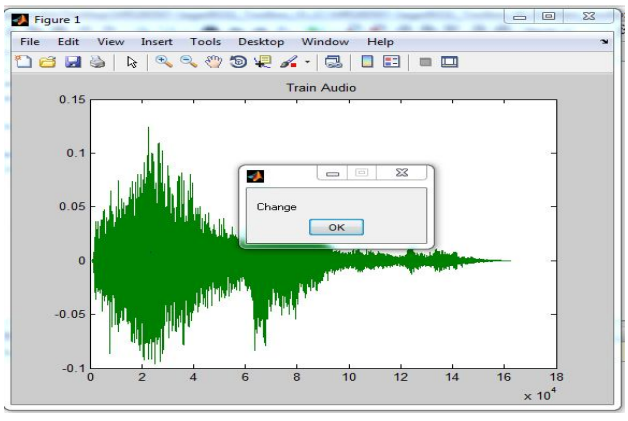

Fig: Classifying audio from video

The video files which were tested during query stage will be compared with the audio files that are trained and stored in the database during feature collection stage, the video files which matches with the audio files are extracted as output.

If none of the video files are matches with the audio files of data base, then the nearest audio file to video, that audio will be generated as output.

V. CONCLUSION

Demonstrating research innovations in a real time environment is valuable for the public to appreciate the work. The research work will publicize the research on audio-visual speech recognition. In addition, if we can make the software reliable enough, it can be used as a module (library) for developing human-computer interface systems with many different components.

Kmeans algorithm is used instead of SVM because outer fittings and inneter fittings in classification handled in kmns, whereas it's not their in SVM, which affects accuracy in classification.

In this proposed system audio of the particular region has been taken and speech processing technique is used to identify speech.

In future work can be concentrated on converting the text file to the audio, so that can create audio into visual speech.

REFERENCES

[1] Gunter, S., &Bunke, H. (2003). Optimizing the number of states, training iterations and Gaussians in an HMM-based handwritten word recognizer. Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. Retrieved November9,2015

[2] Mel Frequency Cepstral Coefficient (MFCC) tutorial. (n.d.). Retrieved November 9, 2015, from

[3] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.,Woodland, P. (2006). The HTK book (Version 3.4 ed.). Cambridge: Entropic Cambridge Research Laboratory.

[4] Noisereduction.(2006).RetrievedNovember10,2015,fromhttp://sound.eti.pg.gda.pl/denoise/noise.html

[5] B.H. Juang, Lawrence R. Rabiner. Automatic Speech Recognition – A Brief History of the Technology Development. Georgia Institute of Technology, Atlanta and Rutgers University and the University of California, Santa Barbara.

[6] H. Silen, E. Helander, and M. Gabbouj, "Prediction of voice aperiodicity based on spectral representations in

HMM speech synthesis," in Proc. INTERSPEECH, 2011, pp. 105–108

[7] T. Le Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in Proc. INTERSPEECH, 2015, pp. 3355– 3359.

[8] Nicolás Morales1, John H. L. Hansen2 and Doorstep T. Toledano1 "MFCC Compensation for improved recognition filtered and band limited speech" Center for Spoken Language Research, University of Colorado at Boulder, Boulder (CO), USA.

[9] M.A.Anusuya ,S.K.Katti "Speech Recognition by Machine: A Review" International journal of computer science and Information Security 2009.

[10] Goranka Zoric, "Automatic Lip Synchr0nization by Speech Signal Analysis," Master. Thesis, Faculty of Electrical Engineering and C0mputing, University of Zagreb, Zagreb, Oct-2005.

[11] README-HTK-AUDIO.(n.d.). Retrieved November 9, 2015, from http://mi.eng.cam.ac.uk/pr0jects/sacti/corp0ra/SACTI-1/utterance-audio/README-HTKAUDIO.TXT.