OPUS-An Android Based Speech to Musical Notes Converter

Avani Dharne,
Department of Computer Engineering,
Savitribai Phule Pune University,
Pune.

Shweta Kulkarni,
Department of Computer Engineering,
Savitribai Phule Pune University,
Pune.

Pooja Shinde,
Department of Computer Engineering,
Savitribai Phule Pune University,
Pune.

Sanjana Dutt,
Department of Computer Engineering,
Savitribai Phule Pune University,
Pune.

***Abstract* - Music in the form of sheet script can have applications in a variety of fields. It is much more easier, compact and feasible to store script sheet rather than an entire acoustic audio.With music transcriptions even processing becomes quite easier. Transcription can be used as a visualization of media players, which during a playback will display the required music score.This paper proposes an Android based application that records the tune sung by the user. The apps uses speech processing and speech synthesis with context dependent acoustic model such as HMM filters for removing the unwanted noise. The intermediate data is further processed to convert it into script music by pitch detection.**

***Keywords*-Android, speech recognition, speech synthesis, HMM, music, pitch detection.**

I. INTRODUCTION

There are times when you just hum a tune and think that it would be great if I could pass this on to someone or just save it for future use. But just recording your voice isn’t a good option because sometimes, it’s our voice that we want to eliminate and just listen to the melody. Musicians when composing a song start from a melody or a tune which then results into a full blown song. The only way to store a melody or a tune is script sheet which is both diificult and tedious to write.Moreover, it cannot be written by just anyone but only by the person who has in-depth knowledge regarding the same.We propose an application that will record your tune, and then convert it into corresponding musical notations so that it can be used further. Musical notations are the best way to store music for music enthusiasts. It makes use of speech recognition and speech synthesis to achieve this.

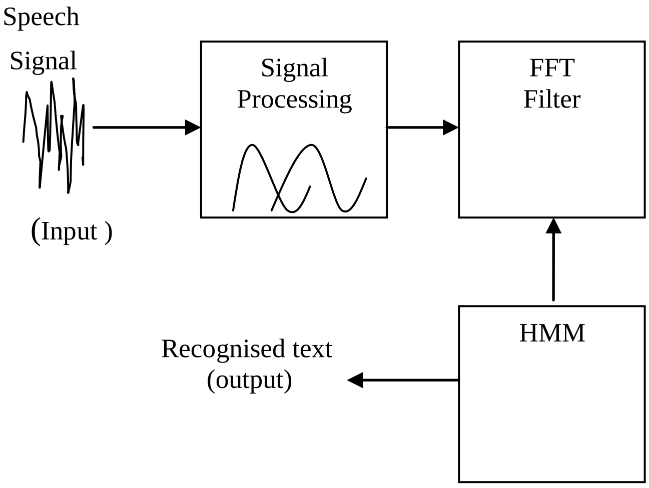

Fig. 1 Block diagram for speech recognition

Speech recognition is the emerging new technology in the field of computer science and artificial intelligence. Speech recognition is basically conversion of spoken speech into text. Android operating system itself consists of an inbuilt microphone that coverts the recorded voice into an equivalent analog signal. This input from the user is also known as utterance. An utterance can be a word, text, song, a sentence or several sentences. The analog signal thus produced is then converted into evenly spaced blocks which then further goes into a FFT filter and is freed of noise and then HMM is applied to recognise the spoken text.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

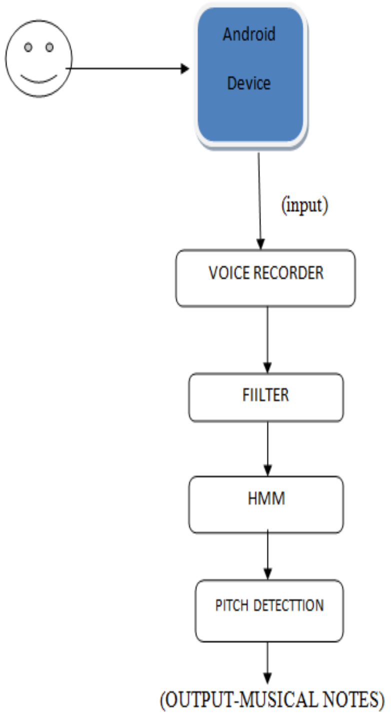

Fig. 2 OPUS architecture

II. THE MARKOV MODEL

Speech recognition has many applications in the industry such as command and control, dictation, transcription of recorded speech, searching audio documents and interactive spoken dialogues. The core of all speech recognition systems consists of a set of statistical models representing the various sounds of the language to be recognized. Since speech has temporal structure and can be encoded as a sequence of spectral vectors spanning the audio frequency range, the Markov Model provides a natural framework for constructing such model.

The Markov model is a type of a bayesian network that can model the time seies data.It contains a list of all the possible states of the system, transition paths beteen those states and rate parameters of those transitions. We will first review the theory of basic markov chains and the proceed over to the hidden markov model.

III. THE MARKOV CHAIN MODEL

The canonical probabilistic model for analysing a temporal or a sequential data is a markov model. Developed by Andrey Markov it follows the basic principle that “The future is independent of the past, given the present.”

Consider a data given by $D=(x_1,x_2,x_3,\ldots..,x_n)$.This data can be modelled by a set a random variables $R=(X_1,X_2,X_3,\ldots.X_n)$ representing the probabilistic markov model .This can be represented as follows :

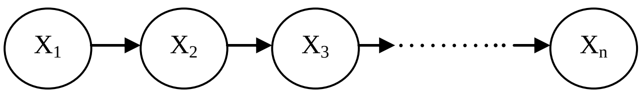

Fig. 2 A Markov chain

This is also known as a markov chain where,
$P(X_t |X_1,X_2,\ldots.,X_{t-1}) = P(X_t|X_{t-1})$ that is,
$P(X_1,X_2,\ldots,X_n)= P(X_1)P(X_2|x_1) P(X_3|X_2)\ldots\ldots P(X_n|X_{n-1})$
Here the next state the variables will transition into depends only on the present state or the present input given. The past states do not have any effect on the present transitions .Thus in every model every X_n state depends on every X_{n-1} state.

IV. THE HIDDEN MARKOV MODEL

A Hidden Markov Model is a statistical Markov model that has some unobserved and hidden states.The diference between a simple markov and a hidden markov model is that in a simple markov model the markov chain(states) is directly visible to the observer and the sequence of states is the output whereas in a hidden markov model the state is not directly visible but the output emitted by that state is visible to the observer. Basically, Hidden Markov Models are finite state automata with transition probabilities and emission probabilities.

HMMs lay at the heart of virtually all modern speech recognition systems and although the basic framework has not changed significantly in the last decade or more, the detailed modeling techniques developed within this framework have evolved to a state of considerable sophistication. The use of HMM in the area of speech recognition is not new.

Automatic speech recognition has a long history of being a difficult problem-the first papers date from about 1950 . During this period, a number of techniques, such as linear-time-scaled word-template matching, dynamic-time-warped word-template matching, linguistically motivated approaches and hidden Markov models (HMM), were used. Of all of the available techniques, HMMs are currently yielding the best performance.

HMM provides with a series of most probable states from a series of observations.A classic use of HMM is in the Hand Writing recognition system where whenever we write some words the responsibility of the underlying system which is HMM is to help to find the next best word. Hmm at the given state holds all the history information upto that state (known as markov property) and then predicts the next state (hidden state) with the best transition probability.

V. DISCRETE MARKOV PROCESS

A Hidden Markov Model H is a quintuple (S, V, π, A, B) where,

- $S=(S_1,S_2,..,S_N)$ is a set of hidden (latent) states that can be observed over a series of time S^T and S always being in one of the N distinct states.
- $V=(V_1,V_2,\ldots V_N)$ is a set of observed states or can be described as the symbols that are emitted due to trasitions in hiddern states.
- π: $S\rightarrow[0,1]=(\pi_1,\pi_2,\ldots.\pi_N)$ is the initial probability distribution on the states which gives the probability of starting in each state.
- $A=(a_{ij})$ where i,j ϵ S is the state transition probability matrix for every transition from one state to another and $\sum(a_{ij})=1$, for each i.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

- B=(b_{jk}) where j ϵV,k ϵ S is the emission probability matrix for every transition from one state to another and ∑(b_{jk}) =1, for each j.

The joint distribution of a sequence of states and observations can be factored in the following way ,

$P(S_{N,}\ V_N)=P(S_1)P(V_1|\ S_1)\ \Pi_{K=2\ TO\ N}P(\ S_k|\ S_{k-1})\ (\ V_k|\ S_k)$.

Let us consider a HMM having 4 visible states and 4 hidden states such that,

S=(S_0,S_1,S_2,S_3) and V=($V_{0,}$$V_1$,$V_2$, V_3) with,

$_{ij}$(a)=

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 0.3 & 0.1 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.7 & 0.1 & 0.1 & 0.1 \end{bmatrix}$$

(b_{jk})=

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.1 \\ 0 & 0.1 & 0.1 & 0.7 \\ 0 & 0.5 & 0.1 & 0.2 \end{bmatrix}$$

The hidden markov can be represented as follows

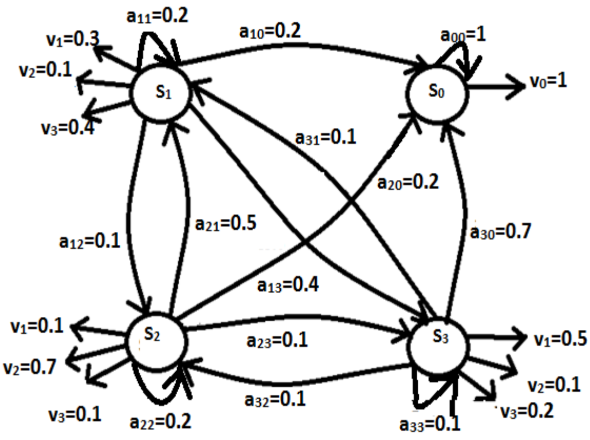

Fig. 3 A hidden markov model

Here in all the figure 3, the system evolves from one state to another emitting symbols with each state having a state transiton probability and an emission transition probability.

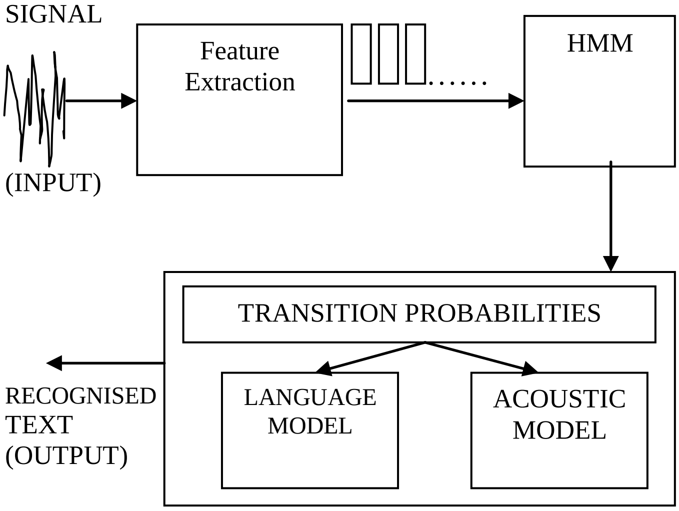

Fig. 4 Hidden markov model in speech recognition.

Here in figure given above the working of a large Vocabulary speech recogniser is illustrated. The input signal received from the microphone is firstly convertred into a fixed sixed acoustic (feature) vectors $W_{1:T}$= ($w_{1,}$$w_2$,….,$w_n$) .This process is knowns as the feature extraction.HMM then tries to find the sequence of words $V_{1:L}$=(V_1,V_2,…V_L) which must have most likely generated W, i.e. it tries to find the transition probabilities for the given W using

$P(W_{1:T,}\ V_{1:L})=P\ (\ W_T|\ W_{T-1})\ (\ V_T|\ W_T)$.

The likelihood of P(V_T| W_T) is determined by the acoustic model whereas the likelihood of P(W_T) is computed b the language model.

VI. FAST FOURIER TRANSFORM

Fast fourier transform is the core of digital signal processing. Fast fourier transform is used to perform fourier analysis and synthesis on a discrete time signal and converts an arbitrary waveform into its sine components..FFT is a very fast computer based algorithm that also gives the DFT (Discrete Fourier Transform). This is the mathematical transition a data goes through when it gets converted from one domain to another,i.e. from time to frequency.FFT not ony work on sounds but also on various forms of continuous signals like radio waves, siesmic data, images ,etc.For performing FFT a given audio input is divided into evenly spaced frames.Each frame contains the same number of fixed samples.To eliminate noise the total energy is measured and also the zero crossings present in the signal are counted as audio input with voice sounds tend to have high volume and low overall fequency i.e. high total energy but lower count of zero crossings while unvoiced sounds tend to have low volume and high overall fequency i.e. low total energy but higher count of zero crossings.Background noise has both low energy and low frequency.This technique helps to eliminate noise and also detect the start and end of the spoken text. FFT also identifies the frequency spectrum of the give waveform which helps in identifying the formants that are the peaks of the frequenct spectrum.As the formants overlap the spectrums that tend to cross theframe boundaries will be missed .

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

VII. PITCH DETECTION

Pitch detection is the fundamental frequency (f0) estimation. Here the basic process is to extract this fundamental frequency from a sound signal which is the lowest frequency component or a partial.

The musical pitch of an audio signal is a perceptual feature relevant only in the context of human listening to the signal. The musical pitch scales can be easily developed only if the frequency and the spectral content were based on the similarity or dissimilarity of the note. Pitch is based on the log of frequency, where the pitch increase about an octave every time the frequency gets doubled. However if the frequency doubles below 1000 Hz it will correspond to a pitch interval slightly less than an octave, if the frequency doubles above 5000 Hz it corresponds to an interval slightly more than a octave.There are many methods for pitch detection that but they mainly operate either in the time.

1. CONCLUSION

Thus, we have studied and applied the theory of Hidden markov model in the context of speech recognition and further processed it with the help of a filter along with pitch recognition to get the desired output in the form of sheet music.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Mr. Sunil Rathod for his immense guidance and constant supervision and Prof. Soumitra Das for his support and encouragement.

REFERENCES

(1) L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, Vol. 77, No. 2, February 1989, http://www.cs.ucsb.edu/~cs281b/papers/HMMs%20-%20Rabiner.pdf

(2) R. L. Cave and L. P. Neuwirth, Hidden Markov models for English, in J. D. Ferguson, editor, Hidden Markov Models for Speech, IDA-CRD, Princeton, NJ, October 1980.

(3) The Brown Corpus of Standard American English, available for download at http://www.cs.toronto.edu/~gpenn/csc401/a1res.html.

(4) T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Speaker Interpolation in HMM-Based Speech Synthesis System," Proc. of EUROSPEECH, vol.5, pp.2523–2526, 1997.

(5) K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Eigenvoices for HMM-based speech synthesis," Proc. of EUROSPEECH, 2002.

(6) A. W. Black, P. Taylor and R. Caley, "The Festival Speech Sythesis System," http://www.festvox.org/festival/.

(7) T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. of ICASSP'92, vol.1, pp.137–140, 1992. [8] http://kt-lab.ics.nitech.ac.jp/˜tokuda/SPTK/ .

(This work is licensed under a Creative Commons Attribution 4.0 International License.)