ISSN:2229-6093
Dr E Chandra et al ,Int.J.Computer Technology & Applications,Vol 3 (4), 1426-1430

An Overview of Speech Recognition and Speech Synthesis Algorithms

Dr.E.Chandra[+], A.Akila[++]
[+]Director, Department of Computer Science,
Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore-32. India.
[++]Assistant Professor, K.G College of Arts and Science, Coimbatore-35.India.
[+]crcspeech@gmail.com , [++] akila.ganesh.a@gmail.com

Abstract

This paper describes about some speech synthesis and speech recognition algorithms and compares their performance based on accuracy and quality. In speech recognition DTW and HMM algorithms are compared with respect to accuracy. Comparative study of CELP and MBROLA algorithm of speech synthesis based on quality is also done.

Keywords: DTW, HMM, CELP, MBROLA, Quality and Accuracy.

1. INTRODUCTION

Speech is a natural mode of communication for people. People learn all the relevant skills during early childhood, without instruction, and they continue to rely on speech communication throughout their lives.

It comes so naturally to the people that they don't realize how complex a phenomenon speech is. The human vocal tract and articulators are biological organs with nonlinear properties, whose operation is not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state.

As a result, vocalizations can vary widely in terms of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed; moreover, during transmission, the irregular speech patterns can be further distorted by background noise and echoes, as well as electrical characteristics (if telephones or other electronic equipment are used). All these sources of variability make speech recognition and speech synthesis, a very complex problem.

2. SPEECH RECOGNITION

Speech recognition is the task of converting any speech signal into its orthographic representation.

2.1 Phases of Speech Recognition

2.1.1 Speech signal. The word spoken is received as sounds and digitized using microphone. The digitized signal is delivered to signal processing unit at a sampling rate not above 8 KHz because sampling rate higher than 8 KHz have less recognition accuracy.

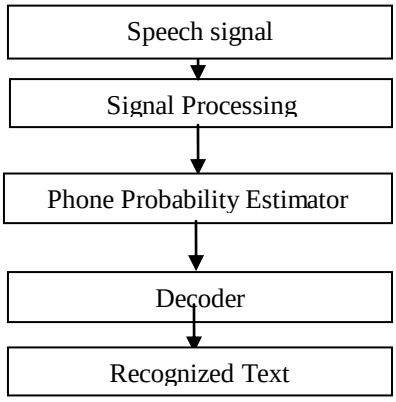

Figure 1: Phases of Speech Recognition

2.1.2 Signal processing. This phase performs feature extraction. Converting linear amplitude signal into spectral like representation [6]. It reduces the data rate of the raw audio input, thereby decreasing the computational load of the fore coming phases.

One type of signal processing that is commonly used in speech recognition systems is Mel Frequency ceptral coefficients (MFCC).

2.1.3 Phone Probability Estimation. This phase estimates the probability that the given features represent a particular sound or combination of sounds in the language [6]. For each sound or combination of sounds, the Phone Probability Estimator outputs a value between 0 and 1 once every time interval. Gaussian Mixture Model is used to estimate the probability.

2.1.4 Decoding. The decoder takes the sequences of estimates of the phone probabilities and compares them against models of every possible utterance in the language. It then outputs the most likely utterance.

The decoder is normally implemented as a search through some famous algorithms like Hidden Markov Model (HMM) and Dynamic Time Warping (DTW).

2.2 Applications of speech recognition

Some of the applications of speech recognition are

1. Data Entry Enhancements in an Electronic Patient Care Report (ePCR).
2. Dictation.
3. Command and Control.
4. Telephony.
5. Wearable.
6. Medical/Disabilities.
7. Embedded Applications.
8. Agricultural application to get farmer queries.

2.3 Speech Decoder Algorithms

2.3.1 Hidden Markov Model Algorithm. Modern general purpose speech recognition systems are generally based on HMM. Hidden Markov Models are standard mathematical technique and their value for modeling processes has been widely recognized from describing models for existing systems to developing test. HMM is a Statistical Model where the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters[1].

Consider a system which may be described at any time as being in one of the state of set of N distinct state, S1, S2, S3, · · · , SN. At regularly time interval system undergoes a change of state (possibly back to same state) according to set of probability associated with the state. Time associated with state change is denoted as t =1, 2 .., and the actual state at time t is denoted as q_t. Calculate probability of occurrence by predecessor

$$\mathbf{a_{ij} = P[q_t = S_i | q_{t-1} = S_j \ ] \ 1 \leq i, j \leq N.}$$

Elements of HMM

1. Number of state N
2. Number of distinct observation symbol per state M, V = V1, V2, · · · VM
3. State transition probability,

$$\mathbf{a_{ij} = P[q_t = S_i | q_{t-1} = S_j \ ] \ 1 \leq i, j \leq N.}$$

4. Observation symbol probability distribution in state j,

$$\mathbf{Bj(K) = P[Vk \text{ at } t | qt = Sj \ ]}$$

5.The initial state distribution $\pi=\pi_i$ where π_i = P[q1 = Si] 1≤ i ≤N[2]

Given appropriate value of N,M,A,B and π, HMM can be used as generator to give an observation sequence

$$\mathbf{O = O1\ O2\ O3 \cdots OT}$$

where O_1, O_2… O_T is observation sequences with time T .

The Three Basic Problem for HMM are

- Evaluation Problem is given observation sequence and model, how to compute probability that observed sequence was produce by the model. Forward Algorithm is used for Evaluation Problem.
- Hidden State Determination (Decoding) is given the observation sequence and model how to choose corresponding state sequence which is optimal in some meaningful sense. Viterbi Algorithm is used for decoding.
- Learning problem is how to adjust the model parameter to optimize model parameter. Baum-Welch Algorithm is used for learning problem.

2.3.2 Dynamic Time Warping (DTW) Algorithm. Dynamic Time Warping algorithm (DTW) is an algorithm that calculates an optimal warping path between two time series[1]. The algorithm calculates both warping path values between the two series and the distance between them. Let (*a*1, *a2*, *an*) and (*b*1, *b*2, *bm*) be two numerical sequences [2]. The length of the two sequences can be different. The algorithm starts with local distances calculation between the elements of the two sequences using different types

of distances. The most frequent used method for distance calculation is the absolute distance between the values of the two elements (Euclidean distance). That results in a matrix of distances having *n* lines and *m* columns of general term:

$$D_{ij}=|a_i-b_j|,\ i=1,n,\ j=1,m$$

Starting with local distances matrix, then the minimal distance matrix between sequences is determined using a dynamic programming algorithm and the following optimization criterion:

$$a_{ij}=d_{ij}+\min(a_{i-1,j-1},a_{i-1,j},a_{i,j-1})$$

where a*ij* is the minimal distance between the subsequences (*a*1,*a*2, ..., *ai*) and (*b*1,b2, ..., *bj*). A warping path is a path through minimal distance matrix from a_{11} element to a_{nm} element consisting of those a_{ij} elements that have formed the a_{nm} distance[2].

Distorsion (in dB)	Recognition accuracy (in %) for DTW	Recognition accuracy (in %) for HMM
Clean	77	92
30 dB	69	73.4
25 dB	45	56.9
20 dB	40	44.1

2.3.3Comparison of HMM and DTW

Table 1

Comparision o DTW and HMM Algorithm based on accuracy

Based on the result produced by various researches, DTW-based system recognition used the MFCC coefficients as basic characteristics vector lead to recognition accuracy of about 77%. HMM-based system recognition presents interesting recognition accuracy with about 92%. In noisy environment, the recognition performances for the two ASR are worse but the pattern recognition using HMM is better than the pattern recognition using DTW.

3.SPEECH SYNTHESIS

The task of speech synthesis is to convert written text (orthographic representation) to speech. The vocabulary should not be restricted for speech synthesis and synthesized speech must be close to natural speech.

3.1 Phases of Speech Synthesis

The general structure of speech synthesis is given here which contains phases like text analysis, Phonetic analysis, prosodic analysis and speech production.

3.1.1 Text Analysis. Input is plain text. Text Analysis tries to understand input text and puts semantic tags into text. In text analysis, Text Normalization is done. Substitution of non text tokens by their text representation is Text tokenization.

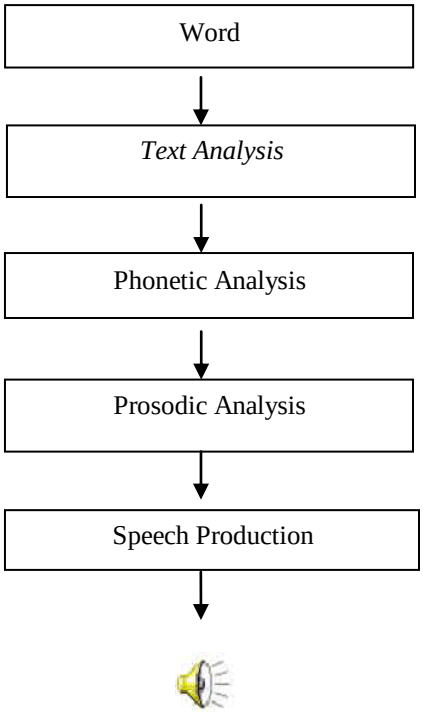

Figure 2: Phases of speech synthesis

3.1.2 Phonetic Analysis. It tries to split text into phonemes. It converts grapheme to phoneme conversion (letter to sound).

3.1.3 Prosodic Analysis. It adds prosodic controls like melody, accent, and pauses to the phoneme string.

3.1.4 Speech Production. It generates speech signal from given string of phonemes and control commands. Some of the frequently used speech production algorithms are MBROLA Algorithm and Code Excited Linear Prediction(CELP).

3.2 Applications of speech synthesis

Some of the applications of speech synthesis

1. For blind -to read books.
2. For educational
3. Telecommunication and multimedia.
4. In warning and alarm.
5. For people with dyslexia.
6. Virtual reality- a news reader in virtual radio station.

7. Agricultural application to give voice response to farmer queries.

3.3 Speech synthesis methods

Speech synthesis methods can be divided into three types.

1. Synthesis based on waveform coding
2. Synthesis based on the analysis-synthesis method
3. Synthesis by rule

3.3.1 Synthesis Based On Waveform Coding. Synthesis based on waveform coding is the method by which short segmental units of human voice, typically words or phrases, are stored and the desired sentence speech is synthesized by selecting and connecting the appropriate units. In this method, the quality of synthesized sentence speech is generally influenced by the quality of the continuity of acoustic features at the connections between units.

3.3.2 Synthesis Based On Analysis-Synthesis Method. In synthesis derived from the analysis-synthesis method, words or phrases of human speech are analyzed based on the speech production model and stored as time sequences of feature parameters. Parameter sequences of appropriate units are connected and supplied to a speech synthesizer to produce the desired spoken message.

3.3.3 Synthesis by Rule. Synthesis by rule is a method for producing any words sentences based on sequences of phonetic syllabic symbols or letters. In this method, feature parameters for fundamental small units of speech such as syllables, phonemes or one-pitch-period speech, are stored and connected by rules. At the same time, prosodic features such as pitch and amplitude are also controlled by rules.

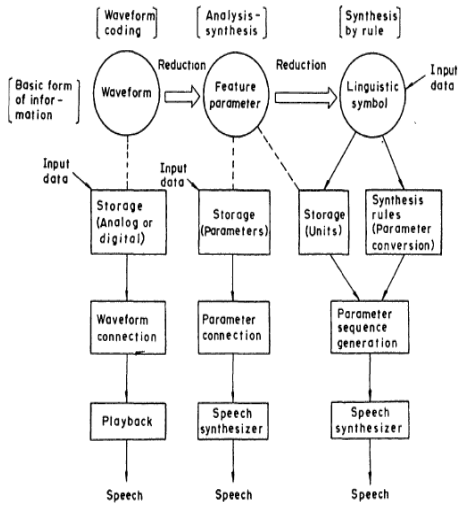

Figure 3: Basic principles of three speech synthesis methods.

3.4 Speech Production Algorithms

3.4.1 Code Excited Linear Prediction (CELP) Algorithm. Code Excited Linear Prediction is an analysis by synthesis procedure introduced by Schroeder and Atal. CELP is the dominant speech synthesis algorithm for bit rates between 4kb/s and 16 kb/s[5].

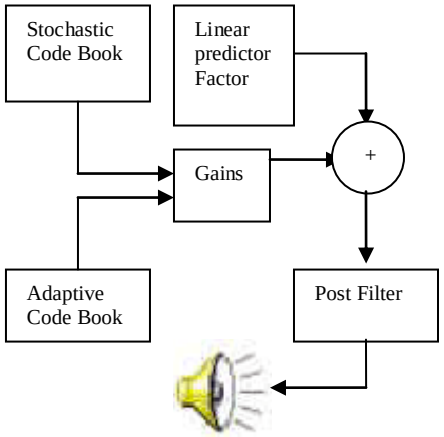

Figure 4: Flow structure in CELP

The adaptive code book contains history of past excitations. The stochastic codebook consisted of independently generated Gaussian random numbers. These codebooks are used to reduce the search complexity. The results got from these searches are gains and added with linear predictor factor to construct a speech[3]. The post filter is used to enhance perceptual quality.

3.4.2 MBROLA Algorithm. Speech synthesis based on the Multiband Resynthesis OverLap-Add (MBROLA) algorithm produces high quality speech without requiring too much effort to design the diphone database, and using a low computational power. The main drawback of this algorithm is the slightly metallic sound or buzziness that can be perceived on voiced segments[4]. The speech quality can be improved by means of an enhanced phase control strategy.

MBROLA speech synthesis uses the PSOLA algorithm applied over a pre-processed speech segments database. This database is obtained by re-synthesizing a natural speech diphone database: first natural speech is coded using a Multiband Excitation (MBE) model, and then decoded with certain modification rules to produce the database used by the PSOLA algorithm. The algorithm is applied pitch synchronously using completely automatic pitch mark generation. This resynthesis algorithm uses a fixed pitch value to avoid pitch mismatches in the PSOLA synthesis stage. It also avoids phase mismatches by using a fixed phase relation between harmonics in every pitch synchronous

frame. This process is applied only over voiced frames.

		Algorithms			
		DTW	**HMM**	**CELP**	**MBROLA**
Parameters	**Accuracy**	77%	92%	-	-
	Quality	Degrades at noisy environment	Degrades at noisy environment	Degrades below 4kbs	Good at all bit rate
	Noisy Environment	Recognition is worse	Recognition is worse	Voice segment has metallic sound	Voice segment has metallic sound

The original spectral envelope in each segment is preserved, so envelope mismatches must be corrected at synthesis time; this can be easily performed by direct time-interpolation between frames, due to the fixed phase relation imposed to the harmonics. These strategies for pitch, phase, and envelope continuity reduce the time employed on the design of the speech units (diphones) database. Moreover the synthesis stage is extremely efficient, reducing to a simple OLA algorithm.

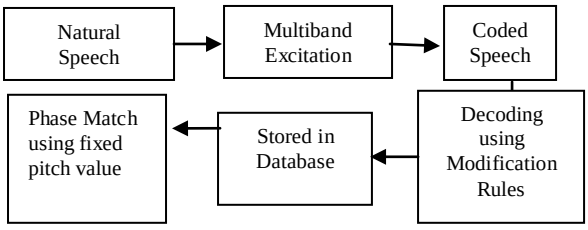

Figure 5: Flow structure in MBROLA

3.4.3 Comparison of CELP and MBROLA Algorithm. Based on the general study of these algorithms, in CELP algorithm the quality degrades below 4kbs due to scarcity of bits to code both excitation and filter parameters. Complexity rate of CELP system implementation employing full codebook searches is very high for commercial applications like agriculture. On the other hand MBROLA algorithm produces high quality speech using low computational power. So MBROLA is much suitable for agricultural application for speech synthesis

.

4 . COMPARISON OF HMM, DTW, CELP AND MBROLA ALGORITHMS

A Comparative study of the existing algorithms of some researchers was made. Based on the study, HMM has more accuracy than DTW. MBROLA has better quality than CELP.

TABLE 2
COMPARISON OF HMM, DTW, CELP AND MBROLA ALGORITHMS

4. CONCLUSION

In this paper, a comparative study of some speech recognition and speech synthesis algorithms was presented. The future work will be to improve the quality rate and accuracy by new hybrid speech synthesis algorithms of CELP and MBROLA and by new hybrid speech recognition algorithm of DTW and HMM.

5. REFERENCES

[1]. Digital speech Processing, Synthesis and Recognition- Sadaoki Furui-Second Edition-Marcedl Ekkerin, Inc.
[2]. Dynamic Programming Algorithms in Speech Recognition - Titus Felix Furtuna - Revista Informatica Economica no. 2(46)/2008.
[3]. LD-CELP Speech coding algorithm- Arun Kumar and Allen Gersho, IEEE signal processing Letters, Vol 4, No 4, April 1997.
[4]. Improving quality in a speech synthesizer based on The MBROLA algorithm by B. Etxebarria, University of the Basque Country, Spain.
[5]. A Comparison of Speech Coding Algorithms ADPCM *vs* CELP by Shannon Wichman, Department of Electrical Engineering, The University of Texas at Dallas ,December 1999.
[6]. Speech Recognition on Vector Architectures by Adam Louis Janin,Univeristy of California, Berkeley, pg 21-27, December 2004.
[7] Rabiner and Jung, "Fundamental of Speech recognition", Pearson Education,©1993.
[8] R. Bellman and S. Dreyfus, "Applied Dynamic Programming", Princeton, NJ, Princeton University Press, 1962.