Vowel-Plosive of English Word Recognition using HMM

Hemakumar G.

Assistant Professor, Department of Computer Science, Govt., College for Women, Mandya.

Abstract — this paper discusses a speech recognition based on spoken English words formed by vowel, diphthong and plosive and it has been developed and experimented for single speaker.

The success rate of recognition of individually uttered words in experiments is excellent and has reached about 98.86 %. The miss rate of about 1.14% was almost only because of false acceptance. In phonemes classification on an average we have reached 85% and miss classification rate was 15%. We have successfully tested all the words formed by vowel followed by vowel-plosive or plosive-vowels or diphthong-plosive and reached high success rate in recognition of the words. All computations are performed in MATLAB and PRAAT software.

Index Terms: - Speech Production and Speech Analysis, frequency, Acoustic-phonetic.

I. INTRODUCTION

Voice recognition system in general very useful in may tasks. Among those very important applications in our everyday life is secure telephony. Voice-based login and voice locks can also use as security key. We can use the voice print of every human being that is why voice recognition (both speech and speaker) plays its significant role in the field of human electronics (humatronics) and its wide applications.

The fundamental task of the acoustic model in speech recognition is to estimate the correct sub-word or phonetic class label for each frame of the acoustic signal. The phoneme can be defined as the smallest phonetic unit in a language that is capable of conveying a distinct in meaning. The task of speech recognition is complicated by the fact that the information relevant to phoneme classification is spread out in both frequency and time, due to mechanical limits of vocal articulators, other co-articulation effect and phonotatic constraints.

In this paper for word recognition, we have created our own database for the words below showed in table 3. It was observed that almost 15% of all misclassified frames are identified as phonemes within the same phonetic group as the correct target. The classifications of phonetic Group which have been used in these experiments are shown in table 1.

II. Speech Production and Speech Analysis:

There are two main sources of speaker-specific characteristics of speech: physical and learned. Vocal tract shape is an important physical distinguishing factor of speech. The vocal tract is generally considered as the speech production organs above the vocal folds. As the acoustic wave passes through the vocal tract, its frequency content (spectrum) is altered by the resonances of the vocal tract. Vocal tract resonances are called formants. Thus, the vocal tract shape can be estimated from the spectral shape (e.g., formant location and spectral tilt) of the voice signal.

Voice verification systems typically use features derived only from the vocal tract. The human vocal mechanism is driven by an excitation source, which also contains speaker-dependent information. The excitation is generated by airflow from the lungs, carried by the wind pipe through the vocal folds. The excitation can be characterized as phonation, whispering, frication, compression, vibration, or a combination of all of them.

Phonated excitation (phonation) occurs when air flow is modulated by the vocal folds. When the vocal folds are closed, pressure builds up underneath them until they blow apart. Then the folds are drawn back together again by their tension, elasticity, and the Bernoulli Effect. This pulsed air steam, arising from the oscillating vocal folds, excites the vocal tract. The frequency of oscillation is called the fundamental frequency, and it depends upon the length, tension, and mass of the vocal folds. Thus, fundamental frequency is another distinguishing characteristic that is physically based.

Compression excitation results from releasing a completely closed and pressurized vocal tract. This results in silence (during pressure accumulation) followed by a short noise burst. If the release is sudden, a stop or plosive is generated. If the release is gradual, an affricate is formed.

The respiratory plays a role in the resonance properties of the vocal system. The trachea is a pipe, typically 12 cm long and 2 cm in diameter, made up of rings of cartilage joined by connective tissue joining the lungs and the larynx. When the vocal folds are in vibration, there are resonances above and below the folds. Sub-glottal resonances are largely dependent upon the properties of the trachea. Because of this physiological dependence, sub-glottal resonances have **speaker-dependent properties**.

III. Signal Preprocessing

In this project, pulse code modulation with a frequency of 8000 Hz, 16-bit mono channel is used. Each word signal contains with a silence region before and after right signal. After resampling, the real signal is segmented and forwarded to the stage of HMM.

IV. Acoustic-Phonetics

The theory of acoustic phonetics is that, which describes the phonetic elements of speech (the basic sounds of the language) and tries to explain how they are acoustically realized in a spoken utterance. These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts. For example, in order to produce a steady vowel sound, the vocal cords need to vibrate (to excite the vocal tract) and the air that propagates through the vocal tract results in sound with natural modes of resonance similar to what occurs in an acoustic tube. Theses natural modes of resonances, called the formants or formant frequencies are manifested as major regions of energy concentration in speech power spectrum. Acoustic phonetics investigates properties like the mean squared amplitude of a waveform, its duration, its fundamental frequency, or other properties of its frequency spectrum, and the relationship of these properties to other branches of phonetics (e.g. articulatory or auditory phonetics), and to abstract linguistic concepts like phones, phrases, or utterances.

V Classification of Vowels, Diphthong and Plosive Speech Sounds:

In the production of vowel sounds the air-current coming from lungs is allowed to go out without any obstruction in the mouth. The diphthong is a gliding monosyllabic speech sound that starts at or near the articulatory position for one vowel and moves to or toward the position for another. The consonant produced by stopping the flow of air at some point and suddenly releasing it are known as plosive sounds. So the sound produces due to the direction in which the tongue moves and variation in the shape of lips result in the change in the shape of the air chamber. It is this particular change which is responsible for the above mention production. Table 1 shows the classification of phonemes used in this project.

Table 1. Classification of Phonemes used in this experiment. [1][3]

Classification	Sub division	Phonemes
Vowels	Front Vowels	i: e ae
	Central Vowels	3 ^ ə
	Back Vowels	a o u ɔ:
Dipthongs	Gliding towards / u /	au, əU
	Gliding towards / i /	ei ai
	Gliding towards / ə /	Iə
Plosive		P, b, t, d, k, g

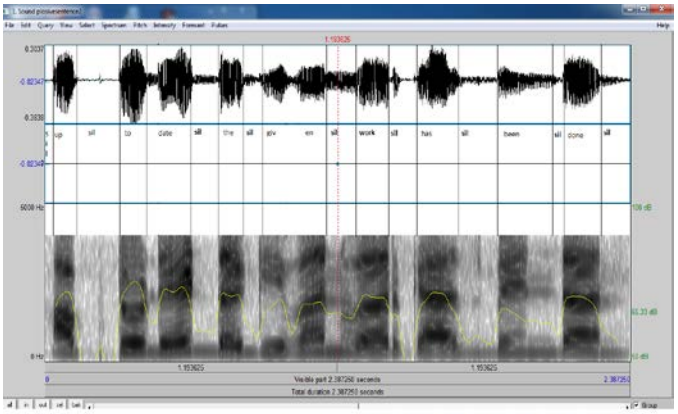

Figure 1 Speech signal and spectrogram of sentence "upto date the given work has been done", adult female voice. The curve inside the spectrogram is showing the intensity of the each phone. sil--> indicating silence region. The dark region in spectrogram shows the Formants. Spech signal acquired at the rate of 8000 Hz, 16-bit, mono channel and analysis done using Praat software.

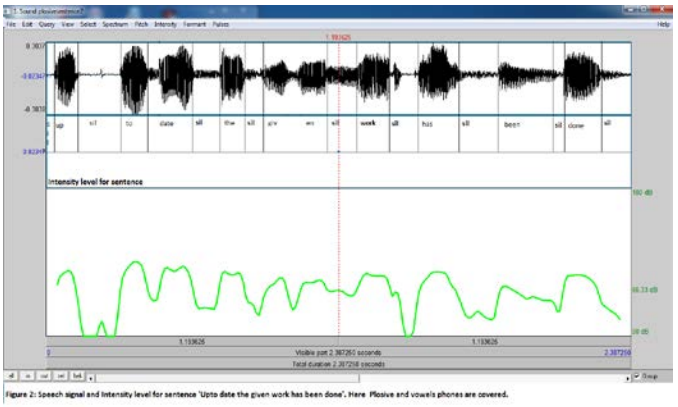

Figure 2: Speech signal and Intensity level for sentence 'Upto date the given work has been done'. Here Plosive and vowels phones are covered.

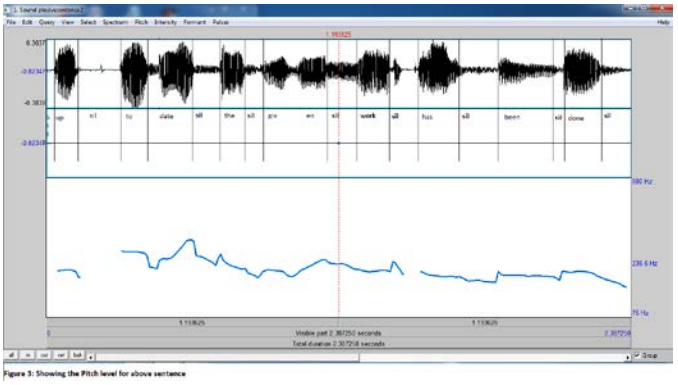

Figure 3: Showing the Pitch level for above sentence

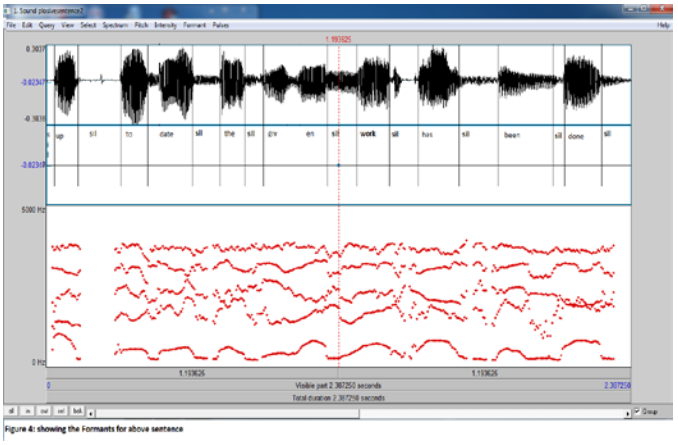

Figure 4: showing the Formants for above sentence

VI. Experimental Results:

Phoneme classification: In this experiment twenty one phonemes are used and classified as Vowels, Diphthongs and Plosives. The phonemes are divided into phonetic classes as shown in table 1, which are based on the distribution of confused phonemes in the confusability matrix. The signal is divided into frames to compute and label the phonemes to the individual cells across the time frequency plane.

In the experiment of recognition of phoneme we have observed that on an average 85% of phonemes are identified correctly and there is 15% of miss classification of phonemes. This misclassification is within the same phoneme group. It is observed that 25% of confusions are still present in an effort to distinguish among vowel groups. Similarly, 20% of confusions are present in an attempt to distinguish among Diphthong groups. However, no such problems are found while distinguishing plosives, due to the clear phone likelihood scoring gap between the plosive. The phone likelihood score of /t/, /d/, /k/, /p/, /b/ and /g/ is in descending order. Table 2 shows the details of success rate and misclassification rate within the phoneme groups.

Table 2: Result of phonemes Recognition.

Phonemes Group	Success rate % of identification with in the group	misclassification / confusions % of phonemes with in the groups
Vowels	75	25
Diphthongs	80	20
Plosive	100	No confusion are found
Average =	85	15

Table 3: Words used for experiment, signal segmentation and corresponding phonemes for the segmented signal as well as HMM states for the phonemes.

Sl. No	**V – P Words**	**Phonemes**	**HMM States**
1.	About	/sil/ - /ə/ - /b/ - /au/ - /t/ /sil/	0→ 1 → 2 → 3 → 4
2.	Add	/sil/ - /ae/ - /d/ - /sil/	0 → 6 → 7
3.	Age	/sil/ - /ei/ - /d/ - /3/ - /sil/	0 → 9 → 7 → 8
4.	Ape	/sil/ - /ei/ - /p/ - /sil/	0 → 9 → 10
5.	At	/sil/ - /ae/ - /t/ - /sil/	0 → 6 → 4
6.	Ate	/sil/ - /ei/ - /t/ - /sil/ /sil/ - /e/ - /t/ - /sil/	0 → 9 → 4 ‘or’ 0 → 12 → 4
7.	Auto	/sil/ - /ɔ:/ - /t/ - /əU/ - /sil/	0 → 11 → 4 → 5
8.	Eat	/sil/ - /i:/ - /t/ - /sil/	0 → 13 → 4
9.	Ebb	/sil/ - /e/ - /b/ - /sil/	0 → 12 → 2
10.	Edge	/sil/ - /e/ - /d/ - /3/ - /sil/	0 → 12 → 7 → 8
11.	Edit	/sil/ - /e/ - /d/ - /I/ - /t/ -/sil/	0 → 12* → 7 → 15 → 4
12.	Egg	/sil/ - /e/ - /g/ - /sil/	0 → 12 → 16
13	Id	/sil/ - /ai/ - /d/ - /i:/ - /sil/	0 → 14* → 7 → 13
14	Idea	/sil/ - /ai/ - /d/ - /Iə/ - /sil/	0 → 14* → 7 → 15
15.	It	/sil/ - /ai/ - /t/ - /i:/ - /sil/	0 → 14* → 4 → 13
16.	Oak	/sil/ - /əU/ - /k/ - /sil/	0 → 5 → 17
17.	Oat	/sil/ - /əU/ - /t/ - /sil/	0 → 5 → 4
18.	Odd	/sil/ - /o/ - /d/ - /sil/	0 → 18 → 7
19.	ok (okay)	/sil/ - /əU/ - /k/ - /ei/ - /sil/	0 → 5* → 17 → 9
20.	Out	/sil/ - /au/ - /t/ - /sil/	0 → 3 → 4
21.	Output	/sil/ - /au/ - /t/ - /p/ - /U/ - /t/ - /sil/	0 → 3 → 4 → 10 → 19 → 4
22.	Up	/sil/ - /^/ - /p/ - /sil/	0 → 22 → 10
23.	Update	/sil/ - /^/ - /p/ - /d/ - /ei/ - /t/ - /sil/	0 → 21 → 10 → 7 → 9 → 4
24.	up-to	/sil/ - /^/ - /p/ - /t/ - / ə/ - /sil/ /sil/ - /^/ - /p/ - /t/ - /əU/ - /sil/	0 → 21 → 10 → 4 → 1* 0 → 21 → 10 → 4 → 5
25.	up-to-date	/sil/ - /^/ - /p/ - /t/ - /ə/ - /d/ - /ei/ - /t/ - /sil/ /sil/-/^/ - /p/ - /t/ - /əU/ - /d/ - /ei/ - /t/ - /sil/	0 → 21 → 10 → 4 → 1* → 7 → 9 → 4 ‘or’ 0 → 21 → 10 → 4 → 5 → 7 → 9 → 4

Note: - * indicates that, this state will be having self-loop and phoneme may be repeated due to the presence of more stress according to speaker dialect.

Table 4: Identification of the Rightly uttered word (Speech Recognition) results		
Sl. No.	Words formed by Vowels-plosive	Recognition % of the right word
1	About	99.8
2	Add	99.2
3	Age	99.2
4	Ape	99.8
5	At	99.8
6	Ate	97.5
7	Auto	98.2
8	Eat	99.3
9	Ebb	99.9
10	Edge	99.2
11	Edit	98.5
12	Egg	99.6
13	Id	97.3
14	Idea	97.2
15	It	96.8
16	Oak	99.8
17	Oat	99.7
18	Odd	99.8
19	Ok	97.2
20	Out	99.9
21	Output	99.8
22	Up	99.6
23	Update	99.8
24	Up-to	97.4
25	Up-to-date	97.2
	Average =	98.86

VII. Conclusion

If phonetic concept is applied for Automatic Speech Recognition, then it makes the software train some other words and it can recognize other words by applying the phonetic to the each word and make a cluster of acoustic phonetic units, which reduces the memory and computation time and at the same time increases the vocabulary of the word recognition.

In this paper, using the observation that phone-level confusions fall more often than not into the same phonemes group as the true target, a phone recognition system was trained to discriminate with in the classes of vowels, Diphthongs, Plosives and comparison between the Vowels and Diphthongs. Here differences in observed probability values are more with vowels-plosive and diphthongs-plosive, therefore phonemes can be identified and labeled to each segments easily.

VIII. Acknowledgement

The author would like to thank Prof. M.R. Nandhan, friends, reviewers and Editorial staff for their efforts in preparation of this paper.

References

1. Lawrence Rabiner, Biing-Hwang Jung, "**Fundamentals of Speech Recognition",** PEARSON EDUCATION (Singapore) PRIVATE LIMITED, Indian Branch, 482 F.I.E Patpargans, Delhi 110092, India, © 1993.

2. Thomas F. Quatieri, **"Discrete-Time Speech Signal Processing Principles & Practice"**, Pearson Education (Singapore) Private. Ltd, Indian Branch, 482 F.I.E Patparganj, Delhi-110092, India, © 2002.

3. T. Balasubramanian, "**A Textbook of English Phonetics for Indian Students",** Published by Rajiv Beri for MACMILLAN INDIA Ltd., 2/10 Ansari Road, Daryaganj, New Delhi 110002, First published 1981 (Reprinted 15 times).

4. Patricia Scanlon, Daniel P.W. Ellis, "Using Broad Phonetic Group Experts for Improved Speech Recoginition", IEEE transaction on Audio, Speech and Language processing, VOL 15, No. 3, March 2007.

5. Khalid Saeed and Mohammad Kheir Nammous, "A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image", IEEE transactions on industrial electronics, VOL 54, No 2, April 2007.

6. Ki-Seung Lee, "Statistical Approach forVoice Personality Transformation", IEEE transaction on Audio, Speech and Language processing, VOL 15, No. 2, Feb-2007.

7. S. Umesh, "Automatic Speech Recognition-Research and standards", Dept of Electrical Enginerring, IIT, Madras, May 7th 2010.

8. Monograph on HMM

9. http://www.audioenglish.net/related/plosive.htm

10.www.audioenglish.net/dictionary/glottal_plosive.htm

11. WIKTOR GONET, "Obstruent Voicing in English and Polish", A Pedagogical Perspective, published by International Journal of English Studies (IJES), vol. 1 (1), 2001, pp. 73-92

Author Biographies : I Hemakumar G. B.Sc, M.Sc, M.Phil. I am having 7 years' teaching experience and 3 years of research experience. I am working for UGC Sponsored minor research project and My interest areas of research are Automatic Speech Recognition, Pattern Recognition and Digital Signal Processing.