SINUSOIDAL ANALYSIS AND MUSIC INSPIRED FILTER BANK FOR TRAINING-FREE SPEECH SEGMENTATION FOR TTS

Ranjani.H.G[1], Ananthakrishnan.G[2], A.G.Ramakrishnan[3]

Dept. of Electrical Engineering, Indian Institute of Science, Bangalore – 560012
INDIA
{ranjani[1], ramkiag[3]} @ee.iisc.ernet.in, gananth[2]@gmail.com

ABSTRACT

The major overhead involved in porting a TTS system from one Indian language to another is in the manual segmentation phase. Manual segmentation of any huge speech corpus is time consuming, tedious and dependent on the person who is segmenting. This correspondence aims at reducing this overhead by automating the segmentation process. A 3-stage, explicit segmentation algorithm is proposed, which uses Quatiere's sinusoidal model of speech in conjunction with a distance function obtained from Bach scale filter bank, to force align the boundaries of phonemes between 2 stop consonants. Preliminary results for Hindi and Tamil sentences show that the misclassified frames (25ms) per sentence or the Frame Error Rate (FER) is 26.7% and 32.6% respectively.

Index Terms— TTS, Phonetic transcription, Sinusoidal model of speech, Bach scale filter bank, Frame Error Rate

1. INTRODUCTION

Many areas of speech research and development take advantage of automatic learning techniques that rely on large segmented and labeled corpora. In speech synthesis, segmentation and annotation of the speech at the phonemic level has become a standard requirement. These corpora are required to train models for recognition as well as to synthesize speech. Text to speech (TTS) systems, in particular, use this synthesis process to create a new concatenation based unit inventory and also for prosody modeling. Hence, the availability of speech data annotated at phonemic level is crucial in the field of speech technology.

TTS systems for Indian languages create standard speech corpora by recording utterances of large number of sentences covering various acoustic and phonetic contexts from a single speaker. The phonetic transcription is available in a TTS corpus as the output of a grapheme to phoneme (G2P) converter. Thus, phonetic labels of speech to be segmented are obtained from the priori knowledge of phonetic transcription. It is only required to align the boundaries of these phonetic transcriptions to their corresponding speech utterances. It is to be noted that this is an automated explicit segmentation technique which has the priori knowledge of the phonetic transcription and hence will not result in "inserted" and "deleted" boundaries.

Automated explicit segmentation using context dependent phone based HMM (CDHMM) gives a good boundary accuracy [8]. Neural network trees with known number of sub word units are used for segmentation [9]. However, the need to develop TTS in multiple languages and the non-availability of large speech corpora for Indian languages are the major constraints which limits the use of segmentation techniques that are training based. The proposed method is a training-free method and gives accuracy comparable with that of methods described in literature.

Segmentation using the Bach scale filter bank has the advantage of being a language independent and training-free algorithm [1], [2]. Section 2 summarizes the above algorithm. Section 3 proposes Quatieri's model for segmenting voiced and unvoiced regions of speech. Section 4 shows the results obtained by combining the above two methods to force align boundaries for segmenting speech at phoneme level. Section 5 concludes the paper and discusses future work.

2. SEGMENTATION USING BACH SCALE FILTER BANK

In this method, speech is treated as a non-stationary signal. A constant Q filter bank motivated by perception of music is formulated. This filter bank has 12 filters in an octave, centers of successive filters separated by a ratio of $2^{(1/12)}$.

Speech signal (Fs = 16 kHz) is passed through this filter bank and the output of these banks at an instant of time is treated as a feature vector. Here, speech is not passed to the filter banks as short segments (frames) as in the quasi-stationary signal and thus, we get feature vectors for every instant of time. Now, the mean of log of the feature vectors in successive 15ms windows is taken and the Euclidean distance between them is calculated. This is referred to as the Euclidean Distance of Mean of Log of feature vectors or EDML [3]. Viewed as a 2 class problem, the distance between the means is a maximum when the feature vectors in the adjacent windows belong to different phoneme classes. Fig 1 shows the contour of the Euclidean distance and the actual boundaries for a part of a speech utterance. Table 1 gives the performance of the method for English, Hindi and Tamil database.

Matched Phoneme Boundary (MPB) is an automated phoneme boundary which falls within 20ms of a manual boundary. If more than one automated segment boundary falls within $\pm$20 ms of a manual boundary or no manual boundary is found within $\pm$40 ms of an automated boundary, then such boundaries are considered to be 'insertions'. Similarly, if no automated boundary is found within $\pm$40 ms of a manual boundary, then it is considered as a 'deletion'.

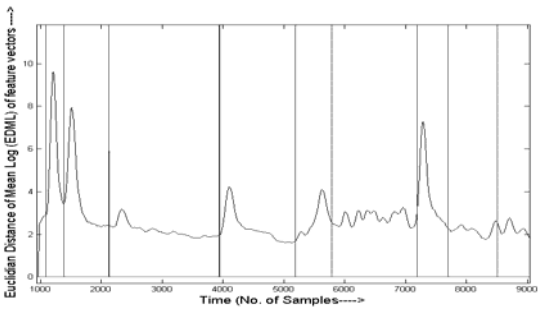

Fig. 1. The plot of EDML against time for a portion of Hindi utterance ("kannarphilm"). The vertical lines denote the actual phone boundaries.

TABLE 1. Segmentation performance of EDML using Bach scale filter-bank for various languages.

Language	%MPB	%Deletions	%Insertions
English (TIMIT)	**82.5**	**22.3**	**18.9**
Hindi	**86.6**	**3.2**	**21.4**
Tamil	**81.9**	**15.3**	**23.7**

3. QUATIERI'S SINUSOIDAL MODEL FOR SPEECH SEGMENTATION

Quatieri proposed a sinusoidal model for speech analysis/synthesis [4]. As per this model, speech is composed of a collection of sinusoidal components of arbitrary amplitudes, frequencies and phases, and it results in an analysis/synthesis system based on explicit sine wave estimation. The first part involves sinusoidal analyzer, in which amplitude, phase and frequency parameters are extracted from the speech signal using STFT. The sinusoidal model of speech signal is expressed as follows:

$$s(n) = \mathrm{Re} \sum_{k=1}^{K} A_k(n) e^{j\theta_k(n)}$$

The next stage involves associating the parameters obtained by the analyzer on a frame-to-frame basis so as to define a set of sine waves that continuously evolve in time. To account for the unequal number of sine components from frame-to-frame, birth and death of these frequency tracks can be noted. Voiced regions give longer-duration frequency tracks and it can be seen that these sinusoids are harmonic in nature while unvoiced regions give short-duration frequency tracks that fluctuate rapidly. This can be explained by Karhunen-Lo`eve analysis, according to which unvoiced signals can only be sufficiently modeled by a very large number of sinusoids. Peaks in unvoiced regions results in many short-duration, rapidly fluctuating tracks. Applying the sinusoidal model of speech for a speech signal, we find that the long-duration frequency tracks indicate start and end of voiced components of speech. This is in-lieu with Quatieri's model and is shown in Fig 2.

Sainath and Hazen [5] use this sinusoidal model for segment-based speech recognizer. The word error rate (WER) using this model was shown to degrade gracefully in presence of noise.

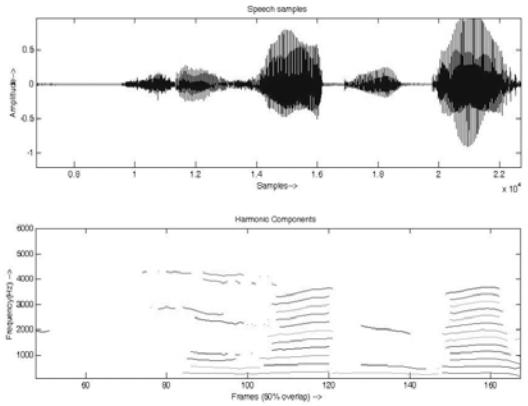

Fig. 2. A portion of TIMIT speech ("She had her dark") waveform and its sinusoidal model

Giving weights to the number of birth and death of sinusoidal components at every frame along with the duration of each track, the voiced and unvoiced regions can be segmented easily.

Results obtained by segmenting speech using sinusoidal model is tabulated in Table 2.

TABLE 2. Comparison of performance of sinusoidal model for different languages

Language	Matched v/uv boundaries (40ms)%	Deleted boundaries%	Inserted boundaries %
English(TIMIT)	**60.2**	**39.9**	**44.7**
Hindi	**68.8**	**31.2**	**50.5**
Tamil	**65.07**	**34.92**	**48.34**

4. FORCED ALIGNMENT USING THE ABOVE ALGORTIHMS

A 3 stage segmentation technique is proposed. The first stage detects stop consonants. The next stage segments the speech waveform between 2 stop consonants into voiced or unvoiced regions using sinusoidal model. The final stage further segments these voiced (or unvoiced) regions at the phonemic level using EDML of Bach scale filter bank. A block diagram of the 3 stage algorithm is given in Fig 3.

The major disadvantage of forcing boundary alignments on the entire speech waveform is that the boundary error gets accumulated. Hence, forcing boundaries for all the phonemes between 2 phoneme classes can be thought of, so that the boundary error occurring at the start of the speech waveform does not propagate to the end of the same. The first stage involves detection of a phoneme class armed with the phonetic transcription, but constrained to be a training-free algorithm.

Fricatives, vowels, nasals, diphthongs, nasal vowels and glides need some stored form of features to be able to detect them. Also, different phonemes in each class require different features. However, stop consonants can be efficiently detected without storing any features as described.

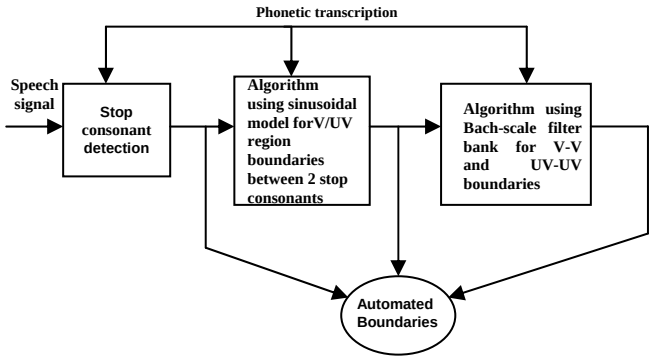

Fig. 3. Block diagram of the 3 stage algorithm for speech segmentation

Predominantly, the first frame (10ms) of any speech sentence is silence. Speech signal is first high pass filtered with cutoff frequency of 400Hz (the voice bar of voiced stops extends till roughly 400 Hz). Now, to detect stop consonants, MFCCs of all the frames in the sentence are taken. Then the Euclidean distance between the first frame and every frame is taken. If this distance drops below a threshold value for a minimum of 3 frames (the minimum duration of a stop consonant can be taken to be roughly 30ms), then the algorithm decides that the region below the threshold approximately contains a stop consonant or a silence region or a combination of both. The frame within this region having the minimum distance from the first silence frame is a sure stop consonant (or silence or both) frame. Experiments give stop consonant detection accuracy of 87% with 20% insertions for 100 sentences from Hindi database.

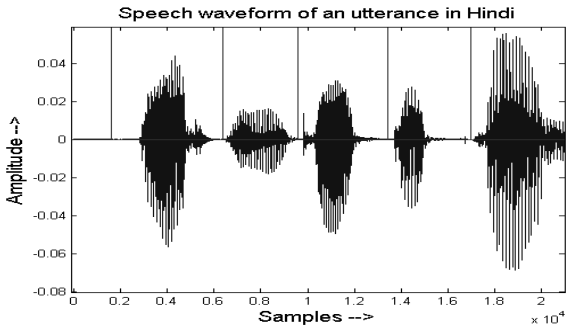

Fig. 4. Portion of a Hindi speech utterance – "or UnkEtatkal". The utterance has 2 silence regions (at the start and end of "or") and 4 stop consonants (of which 2 are contiguous i.e., 'tk'). The vertical lines denote the start of the frame classified as sure stop consonant by the stop consonant detection algorithm.

The number of silence regions involving actual silence regions (between words) and the closure regions of the stop consonants can be known from the phonetic transcription. Using this, the number of silence regions to be detected can be forced. In this case, the equal error rate i.e, the number of insertions is equal to number of deletions, is 11.3% for 100 files in Hindi. For 50 sentences of TIMIT database, the equal error rate is 15%. This is of order, comparable to the stop detection accuracy in literature [6], [7].

The rest of the discussion assumes error-free stop consonant detection and attempts segmentation of individual phonemes between 2 stop consonants.

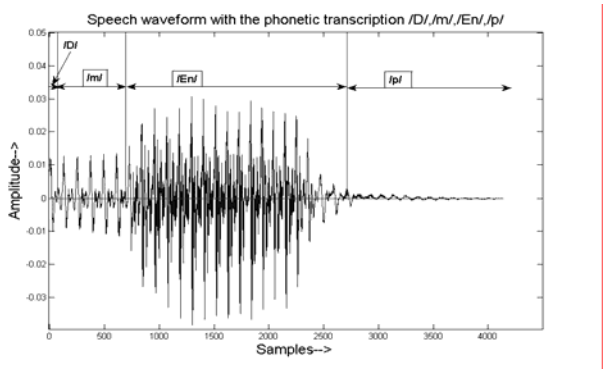

Fig. 5. Waveform of Hindi speech utterance /D/,/m/,/En/,/p/ and its corresponding boundaries.

The phonemes are classified as either voiced or unvoiced phonemes. Unvoiced phonemes are stop consonants and fricatives while all the other phonemes are considered voiced. So, there are regions of unvoiced and voiced speech. Between 2 stop consonants, the number of voiced to unvoiced transitions and vice versa can be counted and the required number of transitions can be detected using Quatieri's speech model as explained in section 3. The output of this stage is the boundaries of voiced and unvoiced regions between 2 stop consonant frames. Fig 6. shows the output of stage 2 for the waveform in Fig 5.

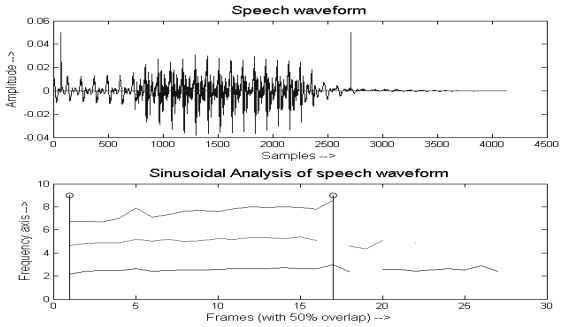

Fig. 6. Sinusoidal analysis of the speech waveform. The vertical lines show the automated boundaries for transitions from unvoiced to voiced transition and vice versa.

The final stage segments further the voiced (unvoiced) regions into the required number of phonemes (for example, the voiced region may have a vowel followed by a glide and then a nasal) . So, the number (N) of required boundaries required is known and the best N such boundaries are chosen from the EDML function from the Bach scale filter bank.

Fig 7. shows the EDML contour for the voiced portion (which contains a nasal followed by a nasal vowel). It can be seen that the distance function peaks at the nasal to nasal vowel transition.

All the above boundaries are combined to form the automated boundaries, shown in Fig 8.

5. RESULTS

The algorithm has been tested on 50 sentences each from TIMIT, Hindi and Tamil databases. The performance of the proposed algorithm i.e., the number of boundaries which lie within a tolerance region of 25ms for 50 sentences of TIMIT database is compared with other methods proposed in literature in Table 4.

Another performance measure for segmenting by force aligning boundaries is the Frame Error Rate (FER). It is the ratio of the number of misclassified frames to the total number of frames in a sentence. Table 5 tabulates the results for the proposed method, assuming error-free detection of stop consonants.

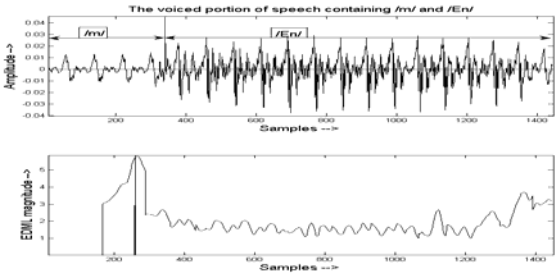

Fig. 7. The voiced portion of speech waveform containing /m/ (nasal) followed by /En/ (nasal vowel) with its corresponding EDML function. The vertical line denotes the automated boundary.

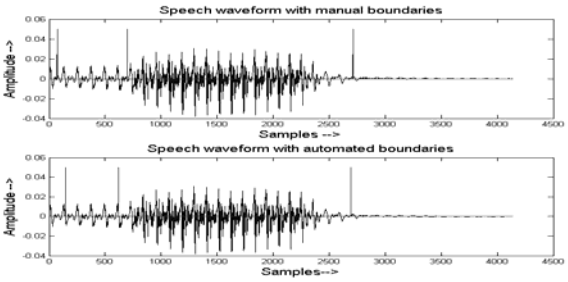

Fig. 8. Speech waveform with the manual boundaries and automated boundaries.

TABLE 4. Comparison of segmentation performances on TIMIT database: Algorithms in the literature as against the proposed one.

Segmentation algorithm	Matched phoneme boundary
NTN [9]	66.6%
HMM [8]	65.7%
Dynamic Programming [8]	70.9%
Proposed method	62.4%

TABLE 5. Frame Error Rate between 2 actual stop consonants of the proposed method for different languages

Language	FER
English (TIMIT)	40.7%
Hindi	26.75%
Tamil	32.6%

6. CONCLUSIONS

The proposed method is language independent and is also training-free. Its accuracy is at least comparable, if not better, to those of the methods that use training algorithms. A final round of manual intervention is required. However, this manual intervention is now less tedious and less time consuming.

It can be seen that the accuracy of sinusoidal model is very low. The sinusoidal analysis can also be viewed as a narrow band spectrogram. Hence, the window size being large, results in a poor time resolution accounting for the low time accuracy of the boundaries. Also, the sinusoidal model, though language independent, needs to be fine tuned for different languages. In other words, the threshold for starting and ending of a sinusoidal track, maximum magnitude increase before birth of a new track and minimum length for a track to exist needs to be changed for effective results. The low accuracy of the method for TIMIT database is justified by the fact that the database comprises different speakers and the model needs to be fine tuned.

Future work can focus on improvement of the stop consonant detection algorithm for higher accuracy and lower insertion. Also, depending on the phoneme transition to be detected, the effect of a variable size window for the Euclidean distance using Bach scale filter bank can be tested. The intuitive concept that detection of some more phoneme classes can improve the accuracy of automated segmentation needs to be verified.

7. REFERENCES

[1] G. Ananthakrishnan, H. G. Ranjani, and A. G. Ramakrishnan, "Language Independent Automated Segmentation of Speech using Bach scale filter-banks", Proc. ICISIP –Dec,2006, pp -115 – 120

[2] G. Ananthakrishnan, H. G. Ranjani, and A. G. Ramakrishnan, "Comparative Study of Filter-Bank Mean-Energy Distance for Automated Segmentation of Speech Signals", Proc ICSCN –Feb,2007, pp 06- 10

[3] Ananthakrishnan G, "Music and Speech Analysis Using the 'Bach' Scale Filter-bank", M.Sc (Engg) thesis, Indian Institute of Science, Apr -2007

[4] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," Proc ASSP Aug, 1986, vol 34, issue 4, pp 744 – 754

[5] T.N. Sainath and T.J. Hazen "A sinusoidal model approach to acoustic landmark detection and Segmentation for robust segment-based speech recognition," Proc ICASSP 2006, pp

[6] F. Malbos, M. Baudry and S. Montresor ," Detection of stop consonants with the wavelet transform", Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, Oct. 1994, pp 612 – 615

[7] P. Niyogia and M. M. Sondhi, "Detecting stop consonants in continuous speech", Proc. JASA Feb, 2002, pp 1063- 1075

[8] Abhinav Sethy, Shrikanth Narayanan, "Refined Speech Segmentation For Concatenative Speech Synthesis", Proc ICSLP-2002, pp 149-152

[9] Sharma, M. Mammone, R., "Automatic speech segmentation using neural tree networks", Neural networks for signal processing, Proc IEEE workshop, Sep 1995, pp 282-290

[10] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition", Pearson education Press, 1993 edition (AT&T).