Speech Synthesis - Automatic Segmentation

Poonam Bansal, Ph.D
Computer Science & Engineering
MSIT, Janakpuri,
New Delhi, India

Amita Pradhan, Ankita Goyal
Computer Science & Engineering
MSIT, Janakpuri,
New Delhi, India

Astha Sharma, Mona Arora
Computer Science & Engineering
MSIT, Janakpuri,
New Delhi, India

ABSTRACT

In this paper, after an a review of the previous work done in this field, the most frequently used approach using Hidden Markov Model (HMM) is used for implementation for phonetic segmentation.

A baseline HMM phonetic segmentation tool is used for segmentation and analysis of speech at phonetic level. The results are approximately same as obtained using manual segmentation.

General Terms

Speech Synthesis, Automatic Segmentation

Keywords

HMM, HTK, Phonetic Segmentation, MFCC, Speech Synthesis, Viterbi

1. INTRODUCTION

Speech segmentation has applications in several fields such as speech and speaker recognition, speech synthesis and speech coding. The segmentation of speech is usually done manually as it is considered to be a source of reliable information. The most precise way to obtain this information is manually. However, manual phonetic labeling and (particularly) segmentation are very costly and require much time and effort. Thus, phonetic segmentation of speech is required.In this work automatic phonetic level segmentation of speech has been accomplished.

Speech segmentation at phonetic level has been implemented using several methods, but the most widely used technique is based on Hidden Markov model (HMM). It is more popular for speech recognition but can also be used for speech synthesis to obtain accurate results [1].

The use of Hidden Markov Models (HMMs) produce a segmentation which, although less precise than a manual segmentation, seems to be precise enough to train the HMMs.

2. HMM - SEGMENTATION OF SPEECH AT PHONETIC LEVEL

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with hidden states. One out of M visible observations is generated randomly by each state.
The following probabilities should be specified for a Hidden Markov model (HMM) :- matrix of transition probabilities A= (aij), aij= P(si | sj) , matrix of observation probabilities B= (bi (vm)), bi(vm)= P(vm | si) and a vector of initial probabilities π= (πi), πi= P(si). Model is represented by M=(A, B, π).

The HMMs were trained using the HTK software and a portion of corpus from TIMIT database. HTK estimates HMM parameters from a set of training utterances. It is very flexible, complete with good documentation.

Steps Followed [1]-[3]

1. *Bootstrapping the model*

The audio file and its sentence-level transcription are input. A wordlist and a dictionary is created using shell script. HDMan is HTK's dictionary management tool. It also outputs a list of phones for which HMMs will be estimated. The model created initially will lack small pause.

Orthographic transcriptions are converted into the HTK label format- mlf (master label file). The same thing is done for the phones in transcripts.

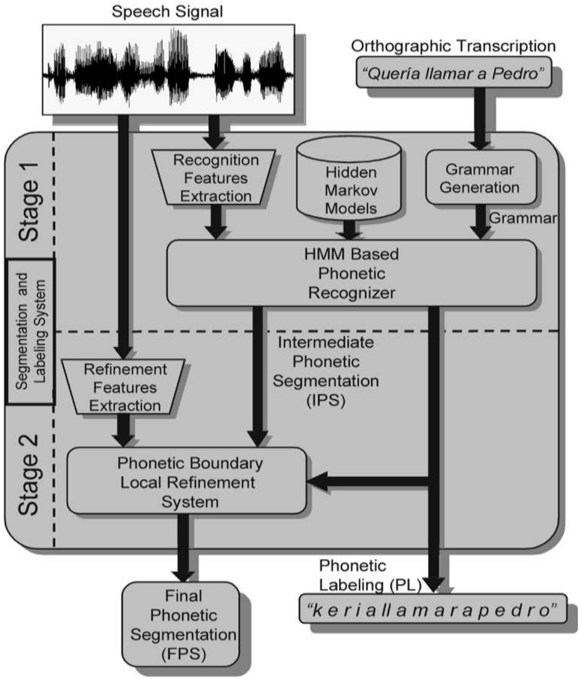
Speech Signal
Orthographic Transcription
"Quería llamar a Pedro"
Stage 1
Recognition Features Extraction
Hidden Markov Models
Grammar Generation
Grammar
HMM Based Phonetic Recognizer
Segmentation and Labeling System
Intermediate Phonetic Segmentation (IPS)
Refinement Features Extraction
Stage 2
Phonetic Boundary Local Refinement System
Phonetic Labeling (PL)
"k e r i a ll a m a r a p e d r o"
Final Phonetic Segmentation (FPS)

Fig 1:Two-stage approach proposed for automatic phonetic segmentation

2. Creating MFCCs

Mel Frequency Cepstral Coefficients (MFCCs),is the standard in speech research. To create the cepstra, which is the raw data used to form HMMs, we use the HTK tool HCopy which takes a single configuration file as input. This configuration file contains information such as sampling rate, pre-emphasis coefficient, window size etc.

3. Initializing the model

To initialize the monophone HMMs called as "flat-start" HMMs since they just take all states to be having the same mean and variance and generates 39 vector values, a 3-state model is used. The model used takes mean to be zero and variance to be one. The HTK tool HCompV is used to compute global mean and variance.

4. Re-estimation

HTK allows re-estimating the flat start monophones using the HTK tool HERest. For this, HMM definitions are created in which each unique phoneme is defined by the prototype. Re-estimation is done thrice to improve the model.

5. Training and segmenting

Now that several versions of the model to be used have been created and trained, it's time to fix a few assumptions that have been made on the way. The first is the two types of "silence" in the corpus- sil, goes at the beginning and end of sentences, and sp, which lacks an HMM. The two should be similar, but not entirely the same, HMM and phones.

To make the models more robust following steps are-

5.i Fixing the silence models

The middle state from the model for sil is copied to build a model for small pause sp. A script-based editor for HMMs, HHed is used.

5.ii Training

Re-estimation is performed twice more with the new model for sp which has been introduced while fixing the silence model.

5.iii Re-aligning data

The point of this re-alignment is to check for alternate pronunciations of words in the dictionary. The generated dictionary may contain multiple pronunciations; at this step, HTK decides which pronunciation is more applicable. Here Viterbi algorithm is implemented using the HVite, HTK tool.

5.iv More training

After the most likely pronunciation has been chosen for each item in the dictionary in the previous step, two more rounds of training are performed using HERest.

5.v Segmenting

We have a sufficient model to obtain time-aligned word and phone transcriptions. The model works by adjusting alignments to maximize the degree to which phones cluster, so HTK should have computed the most likely location of every phone using Viterbi algorithm (within the linear order of a sentence), using the model we've built so far. At this point, there is another possibility for refining the model before outputting the segmentations. One option is to build bi- or tri-phone models. The goal with these types of models is to effectively model co-articulation effects we know to occur pervasively in natural speech. HVite is used once more to output the final segments.

3. COMPARISON TO MANUAL SEGMENTATION

The goal in speech segmentation is not to achieve a perfect phonetic segmentation. Automatic phonetic segmentations are generally evaluated by comparison with segmentations produced manually, which is the most accurate segmentation method known so far, but by no means error-free [6]–[8] .

On comparison of the segmented results obtained using HTK with the manually segmented results using wavesurfer an average % performance of 23.17% is and root mean square % performance of 36.55% is observed.

Considering the boundary mark value of 4ms, % performance for <4ms is 9.42% and for >4ms is 78.14%.

Table 1. Performance Percentage According to boundary values

% performance for <4ms	% performance for >4ms
9.42%	78.14%

Further, on comparison of segmented results using HTK with the segmented samples of TIMIT database and an average % performance of 57.36% and root mean square % performance of 63.67% is observed.

4. CONCLUSION

In this paper, the development of a HMM based segmentation tool for speech synthesis has been described. During the study and implementation of speech segmentation tool, it was observed if more number of states will be used, better alignment and precision is obtained during modelling. Also, i**f** isolated training is done using phonetic transcriptions, better modeling of phone boundary than where manually transcribed training data is available.

Further, another possibility for refining the model before outputting the segmentations is to build bi- or triphone models.

5. REFERENCES

[1] Automatic Phonetic Segmentation, D. T. Toledano, L. A. H. Gómez , Member, IEEE, and L. V. Grande IEEE transactions on speech and audio processing, Vol. 11, No. 6, Nov 2003.

[2] The HTK Book, S. Young, J. Odell, D. Ollason, V.Valtchev, and P. Woodland, Version 2.1: Cambridge University, 1997.

[3] Automatic speech segmentation with HTK, Kyle Gorman, Department of Linguistics,University of Pennsylvania, nstitute for Research in Cognitive Science .

[4] HTK Tutorial,Giampiero Salvi, KTH (Royal Institute of Technology),Dep. of Speech, Music and Hearing, Drottning Kristinas v. 31,SE-100 44, Stockholm, Sweden .

[5] L21, Introduction to Speech Processing | Ricardo Gutierrez Osuna | CSE@TAMU

[6] P. Cosi, D. Falavigna, and M. Omologo, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies," in Proceedings EUROSPEECH, 1991, pp. 693–696.

[7] A. Ljolje, J. Hirschberg, and J. P. H. Van Santen,"Automatic speech segmentation for oncatenative inventory selection," in Progress in Speech Synthesis, J. P. H. Van Santen, Ed: Springer, 1997, pp. 305–311.

[8] A. Ljolje and M. D. Riley, "Automatic segmentation of speech for TTS," in Proceedings EUROSPEECH, 1993, pp. 1445–1448.