Interspeech 2018
2-6 September 2018, Hyderabad

Transcription correction for Indian languages using acoustic signatures

Jeena J Prakash[1], *Golda Brunet Rajan*[2], *Hema A. Murthy*[1]

[1]Indian Institute of Technology, Madras, India
[2]Government College of Engineering, Salem, India

jeena@cse.iitm.ac.in, goldabrunet@gcesalem.edu.in, hema@cse.iitm.ac.in

Abstract

Accurate phonetic transcription of the speech corpus has a significant impact on the performance of speech processing applications especially for low resource languages. Mismatches between the transcriptions and their utterances occur often at phoneme level due to insertion/deletion/substitution errors. This is very common in Indian languages owing to schwa deletion in the context of vowels, and agglutination in the context of consonants.

An attempt is made in this paper to use acoustic cues at the syllable level to remove vowels from the transcription when they are poorly articulated or absent. Hidden Markov model (HMM) based forced Viterbi alignment (FVA) and group delay (GD) based signal processing are employed in tandem to achieve this task. Disagreement between FVA (which produces vowel boundaries based on transcription) and GD boundaries (which uses signal processing cues for syllables) are used to correct the transcription. An increase in likelihood of 0.3% is observed across 3 Indian languages, namely, Gujarati, Telugu and Tamil.

Index Terms: transcription error, group delay correction, forced Viterbi alignment, log-likelihood score, transcription correction

1. Introduction

Robust acoustic models enable better performance of speech systems. In the context of automatic speech recognition (ASR), deep neural network (DNN) based acoustic modeling have been disruptive in that speech based commercial systems have become viable [1]. A DNN based acoustic modeling system requires a huge amount of data for training accurate subword models. Training in ASR systems use word based sentence level transcriptions. The words are converted to phonemes using a standard pronunciation lexicon that assumes a canonical pronunciation for every word based on its articulatory representation [2–4]. Continuous speech often suffers from phoneme insertion/deletion/substitution due to coarticulation [5]. This introduces a mismatch between the speech utterance and corresponding phonetic transcription. Mismatch has a direct impact on the phone models. The primary focus of this paper is to address this issue.

Computationally intensive and data-hungry DNNs are used for building accurate phone models for high resource languages. With large corpus and context dependent phone models, the effect of incorrect transcriptions are overcame since the correct segments average out the effect of bad segments. India has a rich linguistic diversity (22 official languages, 122 major languages and 1599 other languages [6]), but digital resources for even the official languages are scarce. The effect of incorrect transcriptions can affect acoustic modeling significantly [7].

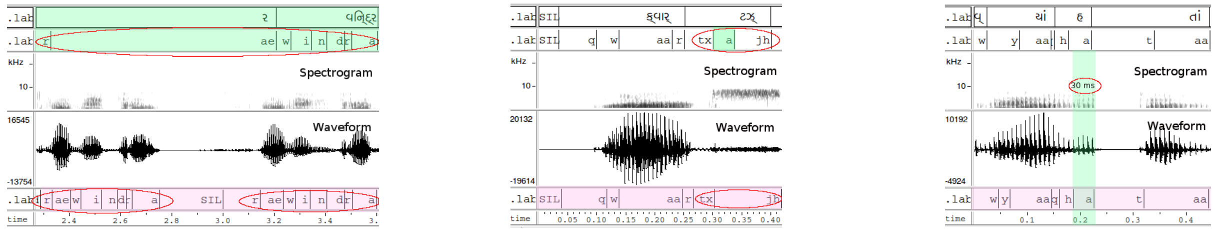
Transcription error mismatches were corrected manually or semiautomatically in Indian languages [8] until recently. This method is tedious, time-consuming, and can be inconsistent across different transcriptors as reported in [9]. An acoustic log likelihood approach was used in [9] to identify possible errors in transcription for Indian languages. In another related work Golda et al. found that even in the TIMIT database (where manual marking is performed carefully) dialectal variations led to different acoustic realisation of words [10]. In this paper, we combine the ideas from [9] and [10] to detect errors in transcription in Indian language datasets.

The database released by Microsoft for the Low Resource Speech Recognition (LRSR) Challenge is used for the study [11]. The database consists of read/conversational speech utterances and the corresponding text transcription of male speakers for three Indian languages, namely Gujarati, Telugu, and Tamil. The text is converted into a sequence of sub-word units (syllables and phones) using a unified parser for Indian languages [2]. Acoustic log-likelihood based approach is used to detect all error-units. Transcription errors are flagged at the phone-level using the approach discussed in [9]. Syllable (represented by C^*VC^*, where C-consonant and V-vowel) being both a production and perception unit [12] are used for identifying missing vowels using the approach discussed in [10]. Both pieces of information are used for removing poorly articulated and missing vowels from the transcription.

A few examples of transcription errors in the Gujarati speech corpus is shown in Figure 1[1]. Figure 1(a) shows an example of insertion error (vowel and consonant insertion) in the Gujarati text transcription. The word $rawindra$ which is present once in the text (marked with a red ellipse in the green panel) is repeated twice in the utterance as indicated in the manual phonetic transcription (marked with red ellipses in the pink panel below the waveform). The frames corresponding to such missing transcripts are assigned to the neighbouring phones, phone *ae* in this case. This paper analyses and identifies such errors automatically for Indian languages. Major causes of vowel insertion/deletion are the schwa deletion rules and poorly articulated vowels. Figures 1(b) and 1(c) show an example of each of these cases. In Figure 1(b), in the segment of speech *qwartz* in a Gujarati utterance, the word *qwartz* is syllabified as *qwar tas* due to wrong parsing, which results in an additional vowel *a* in the phonetic transcription. The error in the phonetic transcription obtained with the parser and the manual phonetic transcription are marked with a red ellipse in the figure. Figure 1(c), the vowel a in the segment *ha taa* is poorly articulated. The duration of *a* is only around 30ms (marked with red ellipse in the figure) while the average duration of a vowel is around 60-70 ms. In such cases, the acoustic features captured differ considerably from that of the expected vowel. This mismatch affects the phone models.

[1] The common label set developed for Indian languages is used for the notation of phonetic transcriptions [13]

10.21437/Interspeech.2018-1188



(a) (b) (c)

Figure 1: *Illustartion of errors in a Gujarati speech corpus. (a) vowel and consonant insertion error (b) vowel insertion error (c) poor vowel articulation. The first and the second panel shows the syllable and phonetic transcriptions by GMM-GD (Gaussian mixture model-group delay) algorithm. The manual phonetic transcription is highlighted in* pink *colour. The errors in transcription (and the poorly articulated vowel) is highlighted in green colour.*

The rest of the paper is organized as follows. Section 2 gives a brief description of the tools used in this study. Section 3 describes an acoustic log-likelihood based approach for detecting transcription errors. Section 4 discusses the group delay driven automatic error correction method. Section 5 analyses the results and Section 6 concludes the work.

2. Overview of techniques used

In order to make the discussion complete, both GD based syllable segmentation and VFA are briefly reviewed.

2.1. Group delay based syllable segmentation

Group delay (GD) segmentation is a speaker and text invariant signal processing technique that finds syllable boundaries in a speech waveform. GD processing is performed on the short-term energy (STE) of the waveform and syllable boundaries appear as peaks in the GD function. The resolution of syllable segmentation using GD is dependent on a parameter called window scale factor (WSF). WSF is inversely proportional to the number of syllables in the utterance. The ability to differentiate close peaks by GD processing is high when WSF is low. The absence of a syllable boundary using GD segmentation in the expected position according to the transcription is an evidence of a missing or poorly articulated vowel in the speech utterance. This can be used as a cue to detect missing vowels in speech utterances.

2.2. Forced Viterbi alignment (FVA)

Hidden Markov model (HMM)-based segmentation is a statistical technique that finds the syllable boundaries where both the text transcription and acoustic models are used to perform a forced alignment. The number of syllable boundaries given by an HMM is the same as the number of vowels in the text. Mismatches in the number of boundaries given by HMM and GD indicates insertion or deletion of vowels.

In this paper, a hybrid HMM-GD segmentation approach is used to build the phone models. [14]. This approach assumes that the transcription has no errors. Vowels and consonants are modelled with 5-state 2-mixtures and 3-state 2-mixtures GMM-HMM (Gaussian mixture model-HMM) models respectively. HTK toolkit is used for HMM training [15]. FVA is performed with these phone models to align each phone in the transcription. Any correction in the transcription leads to FVA performing a realignment, where the frames corresponding to that of the deleted transcription are assigned to neighbouring phones. The acoustic log-likelihood score before and after the transcription correction can be used to flag errors in transcription.

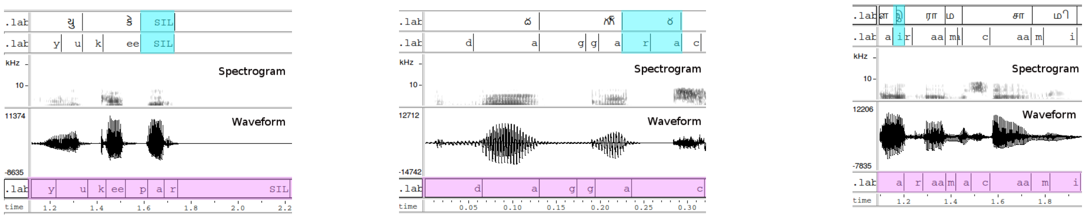
3. Automatic transcription error detection

An approach based on log-likelihood of a phone belonging to different frames of an utterance is devised for the identification of transcription errors. With the given database and lexicon, time-aligned phonetic labels are obtained using HMM-GD segmentation [14]. As mentioned in Section 1, this method does not take into account transcription errors. This could lead to a poor $P(\mathcal{O}|\lambda)$ where $\mathcal{O}$ is the acoustic sequence, λ is the sequence of models that are concatenated based on the transcription. It is important to note that the log-likelihood can be influenced by other factors like segmentation errors[2] as well. Both the cases viz. transcription error and segmentation error contribute to poor acoustic modeling. Hence, a threshold is set on the log-likelihood score to identify the wrong units that could be due to a transcription or segmentation error. The mean μ and variance σ^2 of the log-likelihood score of every phone is computed separately. The phone units whose log-likelihood score lies outside $\mu \pm \mathrm{k}\sigma$ is considered as a candidate for an error.

Figure 2 shows examples of transcription errors in each language used for the study. The phones which are identified as wrong units based on log-likelihood score are highlighted (*blue* colour in the figure) by the proposed algorithm. The wrong units detected with a threshold k at 0.75 (learned in a manner similar to that of [9]) are highlighted in Figure 2. Figure 2(a) shows a Gujarati speech segment, *yukee par*, in which the word *par* which is articulated (shown in the manual transcription - pink panel below the waveform) is missing in the phonetic transcript obtained (highlighted in blue colour in the figure). Due to missing transcript, the corresponding frame gets reassigned to *SIL* (which denotes a silence) and the phonetic alignment becomes incorrect. Figure 2(b) shows an example of deletion error in Telugu language. The syllable *ra*, which is present in the text is absent in the waveform. In Tamil, the word *iraama* is pronounced as *raama* (Figure 2(c)). In this case, the vowel *i* is inserted in the text, which is actually not spoken. Nouns in Tamil that starts with *ra* are generally prefixed by *i*, which is silent in the spoken language. However this can not be considered as a standard rule since a speaker might have articulated the same. Hence it has to be detected from the utterance as to whether *i* has to be discarded or not.

The analysis of the number of error units detected with a threshold on log-likelihood score shows that 21.96%, 26.41%, and 20.06 % of phones are identified as error units in Gujarati, Telugu, and Tamil databases respectively. This includes insertion, deletion, and substitution of vowels and consonants along

[2] The presence of segmentation errors after HMM-GD segmentation is verified manually.



(a)

(b)

(c)

Figure 2: *Illustartion of transcription error detection in (a) Gujarati, (b) Telugu, and (c) Tamil corpus. The first and the second panel shows the syllable and phonetic transcriptions by GMM-GD algorithm. The manual phonetic transcription is highlighted in pink colour. The error units detected by the likelihood based method are highlighted in blue colour.*

with segmentation errors. A further analysis of the errors suggests that among the units identified as candidates for error, 47.21% in Gujarati, 44.83% in Telugu, and 46.14% in Tamil are vowels. Vowel in a syllable corresponds to maximum energy as it forms the nucleus of a syllable, and the duration of the vowel in the syllable is also the largest in a syllable. The next section details the automatic algorithm proposed to correct the transcription for the specific case of vowel deletions in the waveform. Insertion and substitution of vowels, and errors associated with consonants are not dealt with by the algorithm.

4. Group delay directed data-driven transcription correction

Group delay based segmentation gives accurate syllable boundaries when the WSF is chosen such that the syllable rate in the utterance is honoured. HMM FVA gives the exact number of syllable boundaries as obtained from the transcription which is in turn obtained using the pronunciation dictionary or lexicon. We combine GD and HMM FVA to devise an automatic method to correct the transcription when a vowel is poorly articulated or deleted. The location of boundaries given by GD is compared with that of FVA. The absence of any GD boundary near an FVA boundary acts as a cue to the absence of or poorly articulated vowel. However, the number of GD boundaries depends on WSF. With lower values of WSF, GD gives many spurious boundaries which are actually not present. As the WSF is increased, the spurious boundaries reduce. Hence, we start with a very low value of WSF and then increase WSF iteratively. The initial value of WSF as 2 empirically. No syllable boundaries are missed with this value in the initial computation of the syllable boundaries. In every iteration, all the GD boundaries are associated with the closest FVA boundary. An unassociated FVA boundary corresponds to a deletion error. The deleted vowel could be in the FVA syllable under consideration or nearby syllables. In any iteration, whenever an unassociated FVA boundary is obtained, the vowel in the current, previous, and next syllable are removed from the transcription one at a time (basically consider only adjacent syllables for vowel deletion), and the acoustic log-likelihood score is calculated with the original and new transcription in each of the cases. The case with highest likelihood score is chosen as the correct transcription if the duration of the corresponding vowel is greater than 60ms. This is repeated until the association between FVA and GD boundaries stabilises. The GD directed transcription correction algorithm is given in Algorithm 1.

Figures 3 and 4 show a few examples of identification of vowel deletion by GD algorithm. Figure 3 shows a Gujarati utterance *thaar maaq dharpa* (manual transcription is shown in the pink panel). But the transcription given is *thaar maaq thii dharpa* (first and second panel). The word *thii*, that is actually absent in the waveform is inserted in the text. Another error in this segment is the insertion of vowel *a* due to wrong parsing of the word *dharpa* as *dha rap* instead of *dhar pa*. These errors are identified and are highlighted with green colour in the figure. With WSF as 2, GD gives many sharp peaks that correspond to syllable boundaries (last panel) most of which are spurious boundaries. As the WSF is increased, the number of peaks decreases and they get smoothened. With WSF value as 10, the number of boundaries given by GD becomes three, corresponding to the syllables *taar*, *maaq*, and *dhar*. Successive iterations of GD segmentation with different values of WSF provides evidence for vowel deletions. The absence of vowels *ii* and *a* are identified and are marked with red ellipses in the spectrogram. Vowel insertions can also be detected but are not considered as part of this work since GD is not capable of determining the identity of the vowel. The entire sentence transcription, after removing the vowels *ii* and *a*, is given to the HMM FVA algorithm one after the other to confirm the improvement in likelihood score. The consonant *th* is kept as such in the transcription since its duration is short, and GD is incapable of identifying deletion of consonants. Figures 4(a) and 4(b) show similar examples for Telugu and Tamil respectively. The *i*, deleted in the utterance (detected as wrong in Section 3) is identified by GD based data-driven algorithm (WSF 10), the improvement in likelihood score is confirmed with FVA, and is removed from the transcription (Figure 4(a)). The syllable *ku* deleted in the Telugu utterance *la kuu daa* (but given transcription *la ku kuu daa*) is shown in Figure 4(b).

5. Result Analysis

The experiment is performed on 40 hours of male speech corpus of Gujarati, Telugu, and Tamil. The percentage of occurrences of each vowel corrected is shown in Table 1. The percentage of short vowels corrected is higher than that of long vowels. This could be due to the poor articulation of short vowels compared to that of long vowels. Increase in likelihood of the utterance after transcription correction is used as a measure of the correctness of the transcription. Out of 22807, 39141, and 44868 files in Gujarati, Telugu, and Tamil respectively, an improvement in likelihood is observed in 17926 (78.59%), 26948 (60.06%), and 27546 (70.40%) files. Table 2 shows an example of an increase in the cumulative likelihood scores, after transcription correction, of the word *kotxnee* from a Gujarati utterance.

Algorithm 1 Group Delay directed data-driven transcription correction procedure

```
Input:
    1. Waveform W
    2. Viterbi forced-aligned syllables: $(FaB_i, FaE_i), i = 1 \cdots M$ and $FaB$ and $FaE$ are begin and end time, $M$ is the number of syllables given by FVA
    3. Viterbi forced-aligned syllable transcriptions: $FaT_j, j = 1 \cdots M$
Output: Group delay processing directed data-driven corrected transcription
$wsf = 2$
do
    $\{(GdB, GdE)\} = GdSylSeg(W, wsf)$
    $N = |\{(GdB, GdE)\}|$    ▷ N syllable boundaries given by GD
    $\{Cs\} = Associate(\{GdE\}, \{FaE\})$
    for each $c \epsilon Cs$ do
        if $|c| = 0$ then
            $likeli_{orig} = CompLike(W, FaT)$
            for each $syll \epsilon syll_c, syll_{cprev}, syll_{cnext}$ do
                $FaT_{syll} = RemoveSyll(syll, FaT)$
                $likeli_{syll} = CompLike(W, FaT_{syll})$
            end for
        end if
        $(FaT, syll_m) = Min(\{likeli_{syll}\}, likeli_{orig})$
        $flag = FindDur(syll_m, 0.06)$
        if $flag = 1$ then    ▷ Vowel duration $\leq$ than 60ms
            $FaT = CorrTrans(FaT_{syll_{min}})$
        end if
    end for
    $wsf = wsf + 1$
while $N > M$    ▷ Stop iteration when number of FVA boundaries becomes higher than that of GD
```

$GdSylSeg()$ - Gives set of begin and boundaries of all syllables using GD segmentation. $\{(GdB, GdE)\}$ represents the begin and end timestamps of the set of GD syllable boundaries
$Associate()$ - Assign all GD boundaries to the closest FVA boundaries
$CompLike()$ - Compute likelihood score ($likeli_{syll}$) of the given transcription
$RemoveSyll()$ - Correct the transcription by removing the current syllable ($syll_c$) that does not include a group delay boundary, the previous syllable ($syll_{cprev}$) and the next syllable ($syll_{cnext}$) one at a time
$Min()$ - Finds the transcription with maximum likelihood (FaT) and returns the corresponding syllable index ($syll_m$)
$FindDur()$ - Computes the duration of the vowel in the given syllable and returns 1 in $flag$ if the duration is less than 60ms.
$CorrTrans()$ - Returns the corrected transcription FaT.

Table 1: *Percentage of vowels corrected using GD-directed data driven approach. "-" denotes that the corresponding phonetic representation is not used in that language*

Vowels	Percentage corrected		
	Gujarati	Telugu	Tamil
a	10.89	12.21	14.21
aa	7.91	8.27	7.45
ae	21.09	-	-
ai	-	7.67	12.52
au	-	4.82	-
ax	0.52	-	-
e	-	11.23	13.27
ee	0.03	9.01	8.14
ei	-	-	-
i	18.14	12.19	13.96
ii	5.40	6.77	6.19
o	0.03	8.18	8.15
oo	-	9.45	6.10
ou	5.34	-	5.71
u	9.54	14.07	13.38
uu	6.70	8.24	6.94

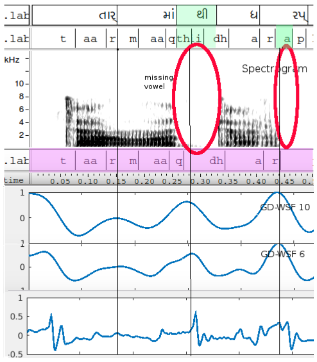


Figure 3: *Example of group delay directed data-driven transcription correction (Gujarati).*

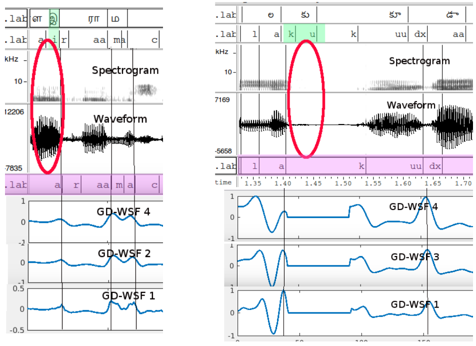


Figure 4: *Example of group delay directed data-driven transcription correction (a) Tamil and (b) Telugu.*

Table 2: *Cumulative likelihood scores of the original versus GD corrected transcription of the word $kotxnee$ from a Gujarati utterance*

Transcription (Original)	Cumulative likelihood	Transcription (GD given)	Cumulative likelihood
k	-1694.43	k	-1694.43
o	-3493.97	o	-3493.97
r	-4481.07	r	-4481.07
tx	-5787.53	tx	-5604.6
ae	-6129.84		
n	-6653.69	n	-6547.26
ee	-8034.43	ee	-7928.00

6. Conclusion

This paper presents an approach where both forced Viterbi alignment and group delay based syllable segmentation can be used in tandem to correct transcription errors. The analysis shows that about 46.06% of the errors flagged are due to vowel deletion. Similar to vowel deletion, other signal processing cues could be used to correct consonants in transcriptions. The increase in likelihood of an utterance does indicate that the deletion of vowels that are either absent or poorly articulated indeed improves phone boundaries. Designing signal processing algorithms that preserve consonant boundaries which last anywhere between 3-10ms is a challenge.

7. Acknowledgement

The authors thank SpeechOcean.com and Microsoft and for using the dataset released for LRSR Challenge 2018 for the study.

8. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] A. Baby, N. L. Nishanthi, A. L. Thomas, and H. A. Murthy, "A unified parser for developing Indian language text to speech synthesizers," in *International Conference on Text, Speech and Dialogue*, Sep 2016, pp. 514–521.

[3] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," 1998.

[4] A. Rudnicky, "Cmu lexicon." [Online]. Available: www.speech.cs.cmu.edu/cgi-bin/cmudict

[5] A. Rechziegel, "On the notions of target and target-undershoot in the phonetics of 11," *Proceedings 22 (l 998)*, vol. 85, p. 95.

[6] Wikipedia, "Languages of india - wikipedia, the free encyclopedia," 2018, [Online; accessed 22-March-2018]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Languages_of_India&oldid=831676831

[7] A. Yuan, N. Ryant, N. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," in *INTERSPEECH, ISCA (2013)*, 2013, p. 2306–2310.

[8] P. Deivapalan, M. Jha, R. Guttikonda, and H. A. Murthy, "Donlabel: an automatic labeling tool for Indian languages," *National conference on communication (NCC)*, pp. 263—-266, 2008.

[9] S. Tanamala, J. J. Prakash, and H. A. Murthy, "A semi-automatic method for transcription error correction for Indian language TTS systems," in *Communications (NCC), 2017 Twenty-third National Conference on*. IEEE, 2017, pp. 1–6.

[10] R. Golda Brunet and A. Hema Murthy, "Transcription correction using group delay processing for continuous speech recognition," *Circuits, Systems, and Signal Processing*, Jul 2017.

[11] Microsoft, "Interspeech 2018 special session: Low Resource Speech Recognition Challenge for Indian Languages," 2018.

[12] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on speech and audio processing*, vol. 9, no. 4, pp. 358–366, 2001.

[13] B. Ramani, S. L. Christina, G. A. Rachel, V. S. Solomi, M. K. Nandwana, A. Prakash, S. A. Shanmugam, R. Krishnan, S. K. Prahalad, K. Samudravijaya *et al.*, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *Eighth ISCA Workshop on Speech Synthesis*, 2013, pp. 311–316.

[14] S. A. Shanmugam and H. Murthy, "A hybrid approach to segmentation of speech using group delay processing and HMM based embedded reestimation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 7334–7338.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book," *Cambridge university engineering department, New Jersy*, 2006.