EUSIPCO 2013 1569746923

ANALYSIS OF VOWEL DELETION IN CONTINUOUS SPEECH

R. Golda Brunet and Hema A Murthy

Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600036, India
e-mail: golda, hema@cse.iitm.ac.in

ABSTRACT

Accurate transcription of the utterances during training is critical for recognition performance. The inherent properties of continuous/spontaneous speech across speakers, such as variation in pronunciation, poorly emphasized or over stressed words/sub-word units can lead to misalignment of the waveform at the sub-word unit level. The misalignment is caused by the deviation of the pronunciation from that defined by the pronunciation lexicon. This leads to insertion/deletion of sub-word units. This is primarily because the transcription is not specific to utterances. In this paper, an attempt is made to correct the transcription at the sub-word unit level using acoustic cues that are available in the waveform. Using sentence-level transcriptions, the transcription of a word is corrected in terms of the phonemes that make up the word. In particular, it is observed that vowels are either inserted or deleted. To support the proposed argument, mispronunciations in continuous speech are substantiated using signal processing and machine learning tools. An automatic data driven annotator exploiting the inferences drawn from the study is used to correct transcription errors. The results show that corrected pronunciations lead to higher likelihood for train utterances in the TIMIT corpus.

Index Terms— speech transcription, pronunciation variability, data driven annotation, acoustic cues and transcription

1. INTRODUCTION

Building a speech recognition system requires an understanding of speech perception. Firstly, it is important to understand the units of speech that are perceivable to the human ear. It has been found in the literature that syllables have a close connection to human speech perception and articulation [1]. The smallest unit of speech production is a syllable. Several psycholinguistic studies have investigated the role of syllable units in speech production and many off-line studies suggest that syllable are the functional units of the speech production process [2]. Greenberg [3] also points out that analysis of pronunciation variation at syllable level is more systematic.

Van Bael et al [4] have demonstrated that the performance of automatic speech recognition systems with automatic transcriptions and manually verified transcriptions are comparable. However, studies on pronunciation modelling [5, 6] shows the importance of orthographic transcription resembling utterances in the context of speech recognition systems. Variability in pronunciation has been accommodated in speech recognition systems by including alternative pronunciations in the pronunciation lexicon. No effort is made to include these pronunciations in the sentence level orthographic transcription of wavefiles, though. In this paper, given the wavefile and its transcription, we propose a modification to the transcription using acoustic cues obtained using signal processing tools. The syllable is used as the unit for study.

The language chosen for the study is English in a clean environment. English being a stress-timed language, the duration between adjacent stressed units is the same. Depending on the distance (in terms of phonemes) between adjacent stressed units, units can either be inserted or deleted. In particular, the units that are deleted or inserted are vowels. In this paper, we refer to any deviation from the form suggested by CMU dictionary [7] as a mispronunciation in continuous/spontaneous speech. For example, the pronunciation of the word ”year” suggested by CMU dictionary [7] is ”y ih r”. Occasionally, there is an extra stress. This might result in the addition of a new vowel ”er” resulting in ”y ih er”. The presence of the additional vowel “er” causes the monosyllabic word of the form CVC to become the bisyllabic word of the form CV VC[1]. This is illustrated in Figure 1. The syllable is defined as a unit that consists of “an onset, rime and coda, with maximum sonority being reached at the rime, while energy decreases towards the onset and coda,” [8]. Figure 1 shows the articulation of three variants of the word *year*. A boundary is drawn where a dip in the energy is observed. This suggests a vowel insertion. This can be seen in the waveform, spectrogram and energy plot. The presence of two prominent peaks in the energy plot suggests the presence of two vowels. Pronunciations can also result in deletion of sub-word units sometimes. For example, Figure 2 illustrates the deletion of the vowel ”ah” in ”ae lah mow niy” (Alimony) . Given that the word consists of four syllables, four humps are expected

[1]C corresponds to a consonant and V corresponds to the vowel.

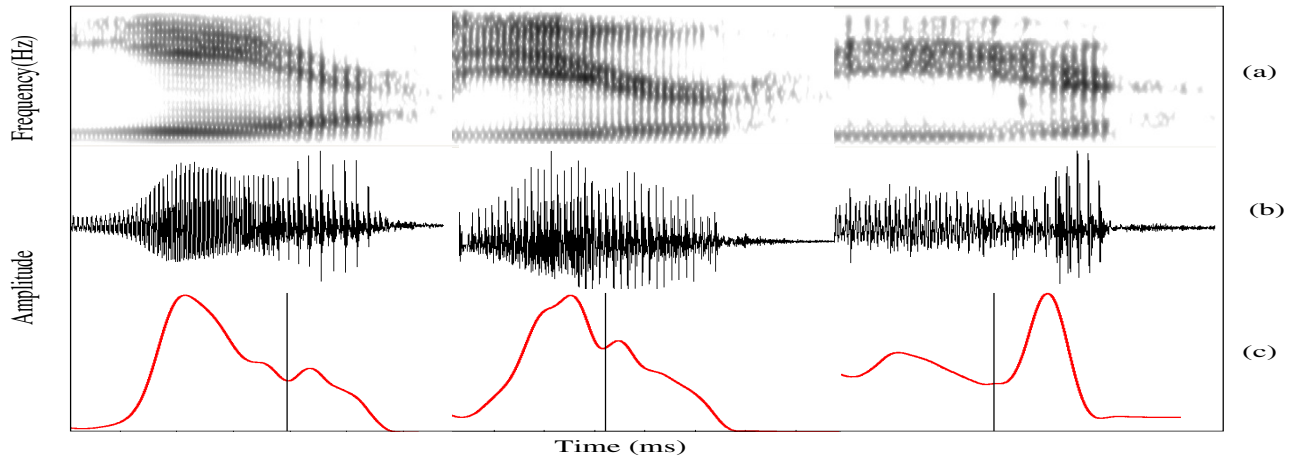

Fig. 1. Addition of new vowel resulting in extra syllable. (a) Spectrogram (b) Waveform (c) Smoothed energy

in the energy contour. Figure 2 shows only three humps. The pronunciation for this utterance is "ael mow niy"[2]. A syllable segment is thus deleted. The vertical bars in all the Figures are possible syllable boundaries.

In this work, an attempt is made to locate these pronunciation variations automatically, given the sentence-level transcriptions. Signal processing and machine learning tools are used to determine the locations of these mispronunciations.

Pronunciation variation has been studied in [9] in the context of syllables. A postmortem analysis on the error statistics of a phone-based context-dependent speech recognizer has been carried out. The observations of the error patterns in the training data have been used to predict the possible errors in the recognized test sentences. Unlike this study, an empirical study [10] shows the impact of pronunciation variation in speech recognition. The manually derived pronunciations are used during recognition and shows performance improvement on a small vocabulary task. Similar to [9], this paper also analyzes pronunciation variability at the syllable. In this paper, we use a group delay based approach in consonance with Viterbi forced alignment and MLP based silence-vowel-consonant detection, to correct the transcription of the training data.

The rest of the paper is organized as follows. Section 2 describes the tools used to derive information about syllables in the continuous speech signal. These cues are then empirically analyzed and observations are made in Section 3. A data driven annotator is designed in Section 4. Section 5 discusses the results and Section 6 concludes the work.

2. OVERVIEW OF THE TOOLS USED

To automate the identification of waveform-transcription misalignment, three tools have been used each giving some information about the syllable. The details of the tools used are explained below.

[2]All pronunciation variations are manually listened to before the claim is made

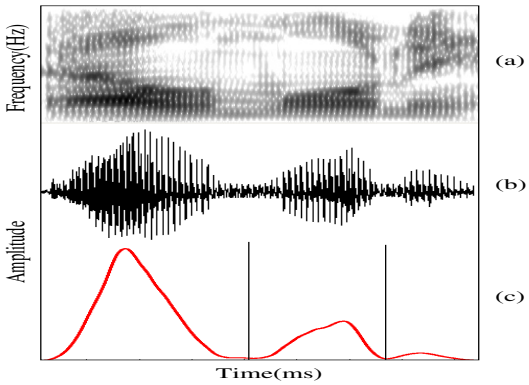

Fig. 2. Deletion of vowel resulting in lose of a syllable. (a) Spectrogram (b) Waveform (c) Smoothed energy

2.1. Group Delay (GD) Segmentation

GD [11, 12] is a signal processing technique to derive the syllable segment boundaries in the speech waveform without the knowledge of the transcription. There is no training involved and only the acoustic cues are exploited to arrive at the syllable boundary information. Figure 3 illustrates the pronunciation of the phrase "carry an oily". The transcription at syllable level for this phrase is "kae riy aen oy liy" having 5 syllables. But the utterance has only 4 syllables "kae riy noy liy" as suggested by the Group delay based boundaries. This is again confirmed by manual listening. This suggests that the transcription is not accurate. If syllable structure is considered as a cue to annotate the waveform, the segment "an" has VC structure while syllable structure in the waveform is CV. This mismatch can be utilized to identify poorly articulated vow-

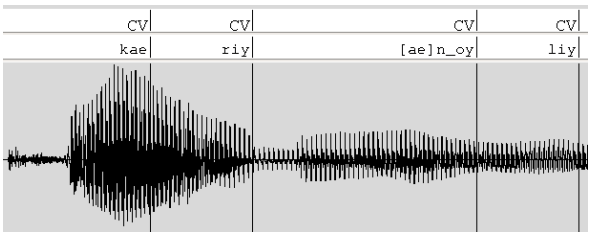

Fig. 3. GD segments for the phrase "carry an oily". [.] indicates missing unit in the utterance.

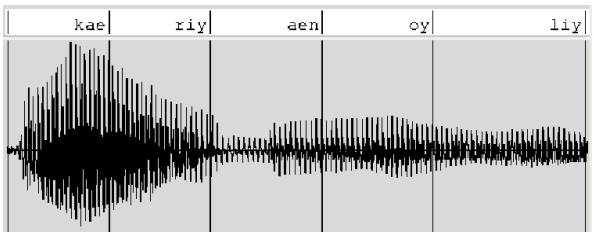

Fig. 4. Viterbi alignment for the phrase "carry an oily".

els. It is straightforward to determine the syllable structure from transcription but to derive the syllable structure from the waveform, a classifier is needed. In the next Section such a classifier is described.

2.2. Silence-Vowel-Consonant (SVC) Classifier

The SVC classifier is a Multi layer perceptron (MLP) based classifier derived by aggregating the output of a MLP based phoneme recognizer (MLP-PR) [13]. MLP-PR is trained with 9-frame context using Modified GD (MODGD) features [14, 15] plus delta coefficients extracted from 100 hours of conversational telephone speech transcribed at phoneme level [16]. There are 45 labels, representing 17 vowels, 27 consonants and 1 silence phoneme. The SVC classifier categorizes the waveform into Silence/Vowel/Consonant at frame level by aggregating the respective outputs of MLP-PR. The consecutive identical outputs are then grouped to give distinct labels in successive blocks. This output is smoothed further merging smaller sized blocks with their neighboring blocks.

2.3. Viterbi Alignment (VA)

A phoneme recognizer is trained and aligned for the best pronunciation in the CMU dictionary [7]. For some utterances, it is observed that number and location of the vowels obtained using GD segmentation and SVC classifier do not match that of the transcription. To correlate the transcription with the segmentation results, VA is used. Each phone is modeled by a 7 state HMM (5 emitting states) employing 3 component Gaussian mixture density functions. HTK [17] is used to train the models with the transcription given in the database. The trained models are then aligned for the best pronunciation in the CMU dictionary [7]. The syllable segment boundaries obtained using VA and the syllable segments from Section 2.1 are correlated. As the transcription used for training and alignment is obtained from the database, the alignment will produce a boundary for every unit in the transcription. For example, the alignment of transcription given by VA for the phrase "carry an oily" in Figure 3 is shown in Figure 4. Although "an" is hardly articulated, a finite number of frames are assigned to the word "an" in the alignment. The group delay segments suggests that perhaps the syllable is missing.

3. EMPIRICAL ANALYSIS AND OBSERVATIONS

The discussion in Section 2 clearly indicates that each tool provides some useful information about the syllables. We now try to combine the results from all the three tools to determine the correct transcription. As mentioned earlier the mismatch most often corresponds to poor emphasis of vowels in syllable segments in the utterance. These observations suggest that such syllable segments correspond to:

- function words as observed by Greenberg [3] (Example: an, and, and so on) or
- word internal biphone syllables (Example: *ae* **lah** *mow niy, per* ***mah*** *nahnt and so on*)

The absence of the vowel is also confirmed by SVC output. Figure 5 provides visual illustration of missing vowels in the utterance. Occasionally the number of GD segments is higher than the number of syllable segments in the transcription. The syllable structure from SVC shows the bisyllable pattern "CV-CV" asserting the presence of extra vowels for a monosyllable in transcription. Interestingly these words have the following characteristics:

- Contiguous vowels (Example: suit - suwt - suw iht) or
- Vowel is followed by "r" (Example: divorced - dih vaorst - dih vao rerst)

It is observed from the empirical study that the presence of consecutive vowels may sometimes contribute to two different vowels in the utterance. For example, the pronunciation of the word "suit" as given in CMU dictionary [7] is "suwt". The syllable structure for this word is "CVC". Occasionally both the consecutive vowels "u" and "i" are uttered by the speaker resulting in a bisyllabic (suw iht) word. The syllable structure of "suit" when it results in a bisyllable pronunciation is "CV VC". The transition region from one vowel to another has a drop in energy resulting in the "C" region in SVC. Thus both GD and SVC confirm the presence of the additional vowel. In the second case, the "r" following the vowel is often uttered

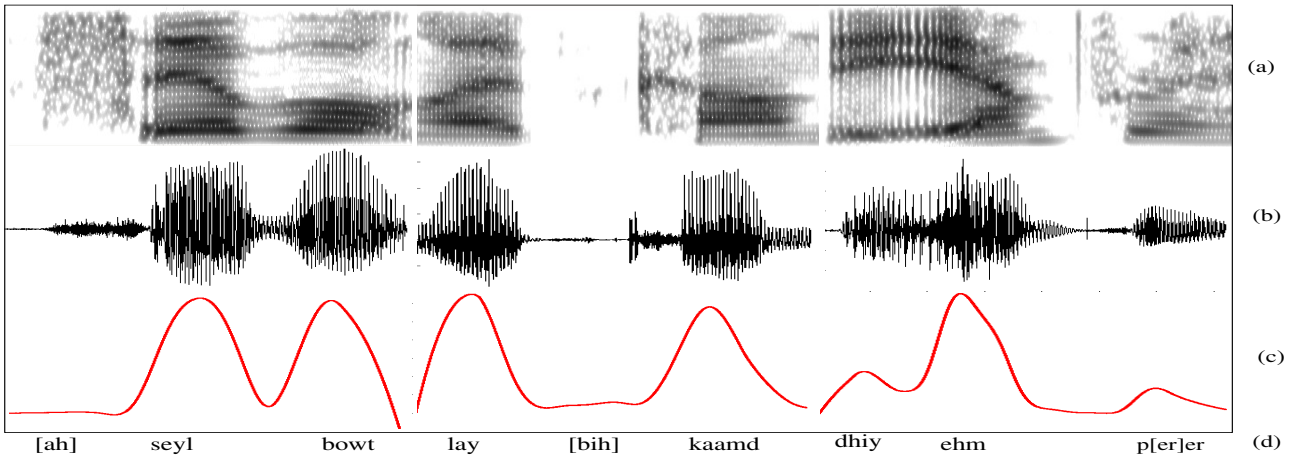

Fig. 5. Poor articulation of vowels. (a) Spectrogram (b) Waveform (c) Smoothed energy (d) syllable transcription. [.] indicates missing unit in the utterance.

as the vowel "er" yielding to a new vowel. The cues from GD and SVC again assert this.

4. AUTOMATIC ANNOTATOR

The observations in Section 3 suggest the development of an automatic annotator that mimics speech production. The VA gives the best pronunciation that matches the utterance of a word. This pronunciation is then syllabified using NIST syllabification software [18] and the syllable boundaries are obtained from VA. An algorithm is now suggested for correcting the transcription:

- Eliminating poorly emphasized vowel:
  - The syllable boundaries from VA and observations from GD and SVC are matched. A poorly articulated vowel or a vowel that is missing will results in missing boundaries in group delay.
  - The absence of the vowel is confirmed using SVC. The transcription is modified to remove the vowel in function words and word internal bi-phone syllables.
- Over stressed vowels:
  - Articulations of vowels not suggested in the pronunciation dictionary, leads to extra syllable segments in GD.
  - The presence of the extra vowel in GD is again confirmed by SVC.
  - However VA does not indicate the presence of the extra vowel since, the lexicon does not suggest the same.
  - These segments are ignored since the phonetic transcription of inserted vowel is unknown.

5. EXPERIMENTAL RESULTS

The experiment is performed on TIMIT corpus [19]. To compare the accuracy of the new transcription, a phone level master label file (MLF) is generated with all training files (4620 files). While generating the MLF, the consonants of poorly emphasized syllables are inserted in appropriate position. This new transcription is then forced aligned with the models used in Section 2.3. The normalized likelihood of each sentence is then compared with the normalized likelihood of the respective forced aligned transcription that comes with the database. Interestingly, the likelihood of the proposed transcription with deleted vowels is high compared to the original transcription. Out of 4620 sentences that are automatically annotated 1402 files(30.35%) have suffered missing of vowels in the utterances. Apart from vowel deletion, silences are also included in appropriate positions indicated by GD segmentation and SVC classifier. The forced alignment of such transcriptions have higher likelihood compared to the original transcription. 1990 sentences which is approximately 43% in 4620 utterances show increased likelihood. Clearly, the increase in likelihood confirms that the acoustic cues obtained using signal processing are indeed correct. The average relative increase in likelihood is 0.3%. Whether such variations, if included in the pronunciation will indeed result in better performance is to be studied. This is because current speech recognition systems are guided quite significantly by the language model. Therefore, subtle increases in likelihood are likely to go unnoticed.

6. CONCLUSION

This paper attempts to identify the misalignment between the transcription and the waveform using cues obtained from signal processing and machine learning tools. With continu-

ous/spontaneous speech it is observed that either vowels are absent or poorly articulated or unnecessary vowels are introduced. The poor emphasis of vowels is common if the syllable is a function word or a word internal bi-phone syllable. Similarly, words with contiguous vowels or words having vowels followed by "r" can give rise to new vowels in the utterance. The missing vowel is removed from the transcription, while the extra vowel is ignored during likelihood computation. It is observed that the corrected transcription results in better likelihood.

7. REFERENCES

[1] Aravind Ganapathiraju, Jonathan Hamaker, Joseph Picone, Mark Ordowski, and George R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE transaction on speech and audio processing*, vol. 9, pp. 358–366, 2001.

[2] Joana Cholin, Niels O. Schiller, and Willem J.M. Levelta, "The preparation of syllables in speech production," *Journal of Memory and Language*, vol. 50, pp. 47–61, 2004.

[3] Steven Greenberg, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," *Elsevier Speech Communication*, vol. 29, pp. 159–176, 1999.

[4] Christophe Van Bael, Lou Boves, Henk van den Heuvel, and Helmer Strik, "Automatic phonetic transcription of large speech corpora," *Computer Speech and Language*, March 2007.

[5] Helmer Strik and Catia Cucchiarini, "Modeling pronunciation variation for asr: A survey of the literature," *Elsevier Speech Communication*, vol. 29, pp. 225–246, 1999.

[6] Kris Demuynck, Tom Laureys, and Steven Gillis, "Automatic generatio of phonetic transcriptions for large speech corpora," *Proceedings of Spoken Language Processing*, September 2002.

[7] "Cmu lexicon," www.speech.cs.cmu.edu/cgi-bin/cmudict.

[8] Xuedong Huang, Alex Acero, and Hsiao Wuen Hon, *Spoken Language Processing*, Prentice Hall Inc., Upper Saddle River, New Jersy, 2001.

[9] Raymond W. M. Ng and Keikichi Hirose, "Syllable: A self-contained unit to model pronunciation variation," *ICASSP*, pp. 4457–4460, 2012.

[10] R. Golda Brunet and Hema A Murthy, "Impact of pronunciation variation in speech recognition," *Proceedings of SPCOM*, July 2012.

[11] T. Nagarajan and Hema A.Murthy, "Group delay based segmentation of spontaneous speech into syllable-like units," *ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition*, pp. 115–118, Apr 2003.

[12] T. Nagarajan, V.Kamakshi Prasad, and Hema A.Murthy, "The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation," *Sixth Biennial Conference of Signal Processing and Communications*, July 2001.

[13] Joel Pinto, G.S.V.S. Sivaram, Mathew Magimai Doss, Hynek Hermansky, and Herve Bourlard, "Analysis of mlp based hierarchical phoneme posterior probability estimator," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 225–241, Feb 2011.

[14] H. A. Murthy and V. R. R. Gadde, "The modified group delay function and its application to phoneme recognition," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 68–71, 2003.

[15] R. M. Hegde, H. A. Murthy, and V. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 190–202, Jan 2007.

[16] R. Padmanabhan, *Studies on voice activity detection and feature diversity for speaker recognition*, PhD Thesis, Indian Institute of Technology, Madras, Aug 2012.

[17] Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, Upper Saddle River, New Jersy, 2006.

[18] W. Fisher, "Nist syllabification software," ftp://jaguar.ncsl.nist.gov/pub/.

[19] Garofolo J. S., Lamel L. F., Fisher W. M., Fiscus J. G., Pallett D. S., and Dahlgren N. L., "Timit acoustic-phonetic continuous speech corpus," 1993.