ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

An alignment matching method to explore pseudosyllable properties across different corpora

Raymond W. M. Ng[1], *Thomas Hain*[1], *Keikichi Hirose*[2]

[1]Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom
[2]Graduate School of Information Science and Technology, The University of Tokyo, Japan

{w.ng,th}@dcs.shef.ac.uk, hirose@gavo.t.u-tokyo.ac.uk

10.21437/Interspeech.2012-262

Abstract

In previous work we have defined a pseudosyllable unit for an English read speech recognition task. In this study we investigate the robustness of extraction of the pseudosyllable units and investigate how such units can be integrated into speech recognition systems. An evaluation method that maps hypothesis phonemes to reference phonemes is proposed. Analysis is performed on pseudosyllables extracted from two different sets of speech data. Mutual information is used to look at the relationship between different pseudosyllabic aspects and the error patterns of hypothesis phonemes. It is shown that the pseudosyllable extraction algorithm is robust and gives units with consistent statistics. Pseudosyllables which have a phone triplet structure tend to have lower insertion error. In temporal regions where pseudosyllables overlap with each other, more insertion errors may occur.

Index Terms: pseudosyllable, error analysis, mutual information, speech recognition

1. Introduction

In automatic speech recognition (ASR), the spoken acoustic signal is decoded into written form. A popular bottom-up approach is to use statistical models to describe the acoustic properties of the smallest meaningful unit in the language, e.g. *phoneme*. From the smallest unit of phonemes, we can work up to word and utterances. Phonotactics and language constraints can be added in order to increase decoding performance.

The addition of high-level constraints such as dictionaries and language models helps to improve ASR performance significantly. On the other hand, the problem of acoustic-phonetic mismatch is seldom tackled directly. Here, we are referring to a problem where a phone is spoken in a manner and place different from its canonical form. If the acoustic properties of speech can be described better, we may be able to save subsequent processing effort.

It is well known that human perception to speech is heavily influenced by the syllable unit [1]. Greenberg pointed to the importance of syllable-level information for understanding of complicated pronunciation variation patterns [2]. In ASR several attempts to incorporate syllable-level information in acoustic modeling were made [3][4][5][6]. Traditional approaches involved the modeling of a large number of syllable-level units, and performance improvement was often marginal. There is a need to evaluate the basis of syllable-level ASR. In a previous study [7], a pseudosyllable unit for English read speech recognition was derived. These units were extracted with little information from the speech recogniser. It was found that phone units in the middle of the pseudosyllables tend to give higher error rate. This suggested the possibility of derivation of some independent confidence measure from syllable-level information to assist in ASR.

Before syllable-level ASR with pseudosyllables can be implemented, there are two questions to be answered. First, the pseudosyllables are extracted with very simple information without considering the spectral properties of speech. The robustness of the extracted unit is an issue. Second, we want to know how pseudosyllable information can help derive the error patterns of phoneme recognition.

In this paper, an extensive test of the pseudosyllable extraction algorithm is carried out. Two different sets of acoustic data are used to test the robustness of the algorithm. Experiments are presented that compare pseudosyllable information and phoneme-class information, in their relationship to phoneme recognition error.

2. Pseudosyllable extraction

This study looks into the syllable structure of spoken language. We use a syllabification algorithm that assumes the least possible knowledge from the phonemic contents in the speech signal. The algorithm assumes that a syllable has a temporal envelope which has a hill-shaped profile. Under this assumption, syllable onsets and offsets are found near the local minima of the temporal envelope, and nuclei at the local peaks. The temporal envelope is represented by the sonorant-band intensity profile. It is obtained by performing waveform rectification and bandpass filtering on the acoustic signal in a specific band known as *voiced band*. By choosing this voiced band with optimal cutoff frequencies, it is expected to exclude the nasal sounds in the low frequency, and the trace in high frequency regions which reflects detailed phonemic identity and pronunciation quality. By referring to previous studies [8][9], cutoff frequencies at 300Hz and 1000Hz are defined for the voiced band.

The assumption of a hill-shaped profile on the temporal envelope guides the search for a syllable. The peaks of the hills are prominent landmarks for syllables [8]. They are first identified by a moving time window. Each detected peak is regarded as the syllable nucleus. The search for syllable onset and offset boundaries is done by evaluating many short segments of the temporal envelope towards both sides of the nucleus. The onset and offset of the hill-shaped profile is characterized by a segment with a concave-upward shape. Thus, syllable boundaries can be indicated by the negative gradient of a segment on the left, and the positive gradient of a segment on the right. To be compatible with other syllabification algorithms, which group the sequences of phones into blocks [10], we define short segments according to speech recogniser output. The resulting

Table 1: *Pseudocode for extracting syllable onset boundary*

```
# All labels indicates the phone position
# For example, c_s is the c_s^th phone from onset
set c_s = peak of s^th syllable
@ k = $c_s − 1;
while ( k > 0 )
    # If k touches the peak of (s − 1)^th syllable
    if ( $k == $c_{s−1} ) then
        @ k = $k + 1
        breakwhile
    endif
    switch (shape(contour(k)))
        case flat : case linear rising :
            @ k = $k − 1
            breaksw
        case upward concave:
            breaksw
        case linear falling: case upward convex:
            @ k = $k + 1
            breakwhile
    endsw
end
set onsetboundary = $k
```

syllable boundaries are aligned with the boundary of hypothesis phonemes. Table 1 shows the pseudocode for finding the syllable onset, a similar mechanism is used to find the syllable offset.

Figure 1 shows four syllables detected by the above algorithm. The temporal span of these syllables is marked by grey horizontal lines. Due to the algorithm implementation, it is possible to have the spans of two neighbouring syllables overlapping with each other. The speech signal and phoneme transcriptions are also included in the figure for reference. There is no consistent definition of a syllable [2][11]. With a proper notation, the suprasegmental unit derived from this algorithm will be referred to as *pseudosyllable*.

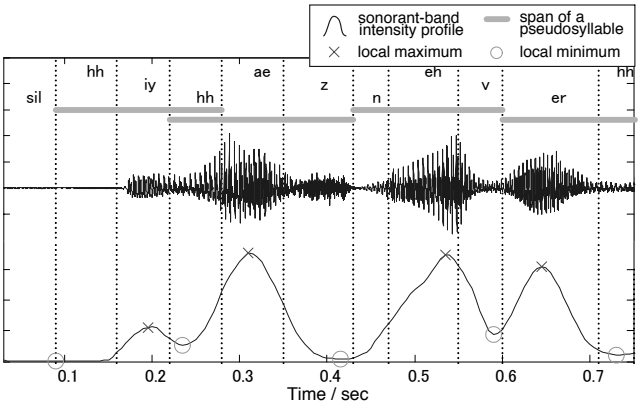

Figure 1: *Syllabification with voiced band intensity profile and alignment information from the speech recogniser*

3. Evaluation method

In this study, we analyse the error rate of phonemes in different locations in a pseudosyllable . Instead of dynamic warping between a sequence of reference and hypothesis phonemes to compare correctness, we match the alignments between hypothesis and reference phonemes to generate a correctness measure.

For this evaluation, a hypothesis phoneme time mark is assigned to every hypothesis phoneme at the mid-point between its onset and offset alignment. It is then associated with the reference phoneme inside which the time mark falls. If there are $N + 1$ hypothesis phonemes associated with a reference phoneme, N insertion errors are marked. The phonemic contents of the reference and the hypothesis are compared to generate a substitution error label. Figure 2 shows both the reference and hypothesis phoneme sequences. Phoneme labels in white fonts under grey shadows indicate insertion errors. Phoneme labels encapsulated in rectangles are substitutions.

Comparing with the typical scoring method based on dynamic programming to map hypothesis phonemes to the reference, the alignment matching method is expected to give lower accuracy and correctness. That is illustrated in Figure 2. A deletion error is indicated for the seventh phoneme "r" in reference because the time mark of the hypothesis does not fall exactly inside the reference alignment.

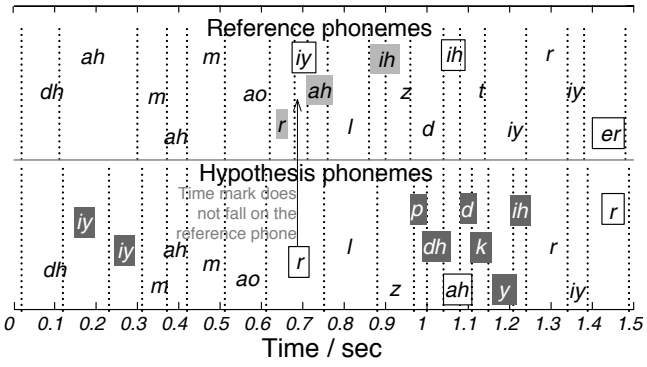

Figure 2: *Evaluation of phoneme recognition results by the alignment matching method*

Unlike conventional error measure which is computed on the reference phonemes, insertion and substitution reported in this paper are calculated with respect to the hypothesis. In summary,

$$\begin{aligned} &\text{No. of Insertion} = \text{Total no. of hypothesis} \times \%\ \text{Insertion},\\ &\text{No. of Substitution} = \text{Total no. of hypothesis} \times \%\ \text{Substitution},\\ &\%\,\text{Insertion}+\%\,\text{Substitution}+\%\,\text{Correct hypothesis}=100\% \end{aligned} \quad (1)$$

Next the relationship between insertion/substitution errors and different pseudosyllable attributes is established by using mutual information. Let E be the binary random variable for *insertion* or *substitution* of a phoneme. F can be a binary feature function which represents an arbitrary *pseudosyllabic aspect*. For example, whether a phoneme is located at the beginning of a pseudosyllable. Mutual information denotes the information of E given by F, and vice versa,

$$I(E;F) = H(E) + H(F) - H(E,F) \quad (2)$$
$$\text{where } H(X) = \sum_{x \in \mathcal{X}} p(x) \times log\frac{1}{p(x)}$$

High mutual information indicates that a particular feature value contains more information for inducing an error feature, and vice versa. This happens when the cross entropy, $H(E,F)$, is small. Another condition for high mutual information is high

entropy $H(E)$ and $H(F)$. This requires the feature values having a relatively even distribution. In another words, the population between in-class and out-of-class samples has to be balanced.

4. Experimental data and analysis method

4.1. Data

The pseudosyllable extraction algorithm is applied to two different English speech databases to test its robustness. The first set is English read speech recorded in clean environment from the TIMIT database [12]. Each utterance is phonemically balanced and is about 3 seconds long. We extend from the 192 core test set utterances we used in our previous experiment to the full testing set with 1344 testing utterances [7]. We refer to this set as **TIMIT** hereinafter.

The second data set is evaluation data from NIST evaluations on meeting data in 2007 (rt07seval) [13]. Speech data are in 35 ten-minute excerpts recorded by independent head microphones from 4 meetings. Each speaker has a separate channel and manual segmentation is applied to give 4522 testing utterances. Each utterance is 30 seconds long on average. We refer to this set as **rt07** hereinafter.

4.2. ASR system

Different phoneme models are trained for the two data sets. For **TIMIT**, Mel-frequency cepstral coefficients (MFCCs) are used without any normalisation. The system for **rt07** is more sophisticated [14]. It uses perceptual linear prediction (PLP) adaptation, cepstral mean and variance normalisation per speaker and vocal tract length normalisation. Hidden markov models are trained for the English phonemes in a conventional context-dependent triphone tied state structure. The system for **TIMIT** follows the maximum likelihood (ML) training criterion. The system for **rt07** uses Maximum Likelihood Linear Regression, MPE training and hLDA.

In the decoding stage, dynamic programming implemented by Viterbi algorithm using an all-phone network is applied to generate hypothesis phoneme sequences for the test data. Different phoneme insertion penalties are tried for **TIMIT** and **rt07** and an optimal values of -20 and -15 are set respectively. The reference phonemes are generated by forced alignments. **TIMIT** and **rt07** data use their own corresponding acoustic models for the generation of forced alignments.

4.3. Analysis method

To analyse the robustness of the pseudosyllable extraction algorithm, the properties of pseudosyllables across **TIMIT** and **rt07** are compared. Then, the relationship between the pseudosyllabic aspects of a phoneme and its recognition error is studied. The possible pseudosyllabic aspects of a phoneme include the position of the phoneme in its corresponding pseudosyllable, the length of the corresponding pseudosyllables, etc. The relationship between phonemic type and its recognition error is also studied.

5. Statistics of the pseudosyllable

Table 2 shows pseudosyllable statistics extracted from the two test conditions. The total number of pseudosyllables in **TIMIT** and **rt07** is 16k and 44k respectively. The average number of phones in a pseudosyllable is from 2.5 to 2.7. Most of the pseudosyllables are phone pairs, triplets and quadruplets. They add up to 84% of the pseudosyllables in **TIMIT** and 76% in **rt07**. The proportion of single-phone pseudosyllable in the **rt07** (11%) is higher than that in **TIMIT** (7%). This is because some speech segments are short. It is also believed the abundance in single-phone pseudosyllable reflects the nature of conversational speech.

The average duration of syllables is around 270ms. We can see the comparable duration patterns across two sets, while the **rt07** set gives higher variance than the **TIMIT** set.

We take a closer look to the pseudosyllables comprising three phones. We break down all English phonemes into three categories - consonants (C), vowels (V), and semi-vowel (L). There are only three major constructions among many others which give a frequency higher than 5%. They are CVC (34.84% for **rt07**, 33.66% for **TIMIT**), LVC (25.21% for **rt07**, 25.91% for **TIMIT**) and CVL (9.52% for **rt07**, 11.08% for **TIMIT**).

The above statistics indicate that the pseudosyllables extracted from speech across two very different corpora (**TIMIT** and **rt07**) share a consistent statistic.

Table 2: *Statistics of the pseudosyllables in two corpora*

Statistics	**TIMIT**	**rt07**
Total number of pseudosyllables:	16386	44347
Statistics per pseudosyllable:		
Avg. no. of hypothesis phonemes	2.6	2.5
Avg. no. of reference phonemes	2.7	2.6
Average duration	269ms $\pm$125ms	276ms $\pm$178ms

6. Special error patterns

Table 3 shows the mutual information between the insertion/substitution error and the length of the pseudosyllables in which a phoneme resides. High mutual information is found between insertion and $F = 3$. That indicates a large difference on the insertion in a three-phone pseudosyllables (9.0% for **rt07**; 3.8% for **TIMIT**) and that in other pseudosyllables (15.6% for **rt07**; 7.7% for **TIMIT**).

Table 3: *Mutual information $I(E;F)$ between error types and the length of the pseudosyllable in which a phoneme resides*

F:Pseudosyllabic aspects	E:Insertion		E:Substitution	
	TIMIT	**rt07**	**TIMIT**	**rt07**
(Length of pseudosyllable in which a phoneme resides)				
1	10.8	28.6	10.9	40.2
2	125.5	67.7	81.0	74.9
3	**361.6**	**517.2**	30.3	0.1
4	56.4	27.6	19.3	2.5
5	126.1	41.0	24.5	32.0

All values are scaled by $\times 10^5$

Table 4 shows the mutual information connecting different phonemic classes and the errors. The largest amount of information is found in the phoneme classes of vowel and semivowel in **TIMIT** which give substitution errors. Vowels and semivowels occupy 38% and 24% of the phones. While the average substitution rate for a phone is 25.6%, vowels have a higher substitution rate at 32.6% and semivowels have a lower one at 14.2%. The distinctive error patterns for vowels and semi-vowels are less salient in **rt07**, where the corresponding substitution rate is 29.3% and 21.4%.

Table 4: *Mutual information $I(E;F)$ between error types and different classes of a phoneme*

F:Pseudosyllabic aspects	E:Insertion		E:Substitution	
	TIMIT	**rt07**	**TIMIT**	**rt07**
(Phoneme class)				
Consonant	34.3	197.4	206.1	26.5
Vowel	3.0	113.9	**1363.4**	221.5
Semi-vowel	24.1	18.6	**791.4**	137.2
(Phoneme position in the pseudosyllable)				
Onset (o)	85.6	185.3	12.1	0.0
Nucleus (n)	32.0	61.3	90.5	0.2
Coda (c)	100.4	176.3	7.4	13.3
overlap (c-o)	4.5	1.38	137.8	0.0
overlap (n-n)	62.5	**565.0**	0.36	18.7
overlap (c-n/n-o)	**366.8**	**859.5**	53.6	15.0

All values are scaled by $\times 10^5$

Table 5: *Phoneme recognition errors in different classes*

F:Pseudosyllabic aspects	Frequency	E:Insertion		E:Substitution	
		TIMIT	**rt07**	**TIMIT**	**rt07**
Consonant	38%	7.4%	16.8%	20.3%	24.4%
Vowel	38-40%	5.2%	10.5%	32.6%	29.3%
Semi-vowel	22-24%	6.4%	12.8%	14.2%	21.4%
Onset (o)	18-20%	5.4%	10.0%	24.9%	25.5%
Nucleus (n)	34-35%	8.1%	12.6%	26.0%	25.3%
Coda (c)	20-21%	5.1%	11.0%	23.2%	27.4%
overlap (c-o)	10-13%	1.3%	9.1%	17.8%	25.5%
overlap (n-n)	2-3%	14.2%	40.7%	22.8%	20.5%
overlap (c-n/n-o)	3-4%	22.7%	38.5%	14.2%	21.6%

Finally more pseudosyllabic aspects are investigated. **Onset**(o) and **coda**(c) are the first and last phoneme in the pseudosyllable. Phonemes residing elsewhere in the pseudosyllable are defined as **nucleus**(n). Recall that pseudosyllables may overlap with each other. Three cases of overlap (**c-o**, **n-n** and **c-n/n-o**) are considered. For instance, (overlap c-o) refers to coda-onset overlaps. It accounts for 12% of the phonemes in **TIMIT** and 11% in **rt07**. The insertion and substitution errors do not show a special trend. However, it is abnormal to have a phoneme when there are overlaps between two nuclei (overlap n-n) or between nucleus and boundary (overlap c-n/n-o). These phonemes account for 6% in **TIMIT** and 7% in **rt07**. In the two bottom rows in Table 5, a higher insertion is noted to these phonemes. Figure 3 shows the plot of temporal envelope of these phonemes. It can be noticed they are hypothesis phonemes either on a flat contour, or on a complicated profile where the assumption of a single hill-shape profile is broken.

7. Conclusions

In this study, the robustness of the pseudosyllable extraction algorithm is illustrated by its application to two different datasets. From various comparison between different phonemic types and their errors in ASR, it is shown that pseudosyllables which have a phone triplet structure tends to have lower insertion. Pseudosyllables which overlap with their neighbours are places where more insertion errors may occur. The nature of the extracted pseudosyllable units is reflected in the statistics. The analysis framework could be reiterated, such that the relationship between phoneme recognition error conditioned on different aspects of the pseduosyllables can be revealed.

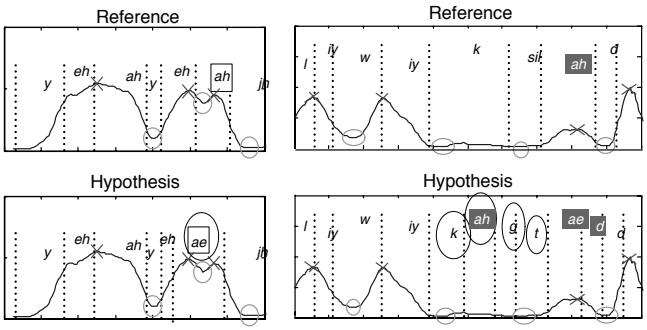

Figure 3: *Phonemes with pseudosyllable overlap : (Left) Two units overlap at /ae/; (Right) The unit in the left extends towards /t/ and that in right extends towards /k/*

8. References

[1] O. Fujimura, "Syllable as a unit of speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 82–87, February 1975.

[2] S. Greenberg, "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2-4, pp. 159–176, November 1999.

[3] 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, May 2001.

[4] H. Wu and X. Wu, "Context dependent syllable acoustic model for continuous Chinese speech recognition," in *Proc. Interspeech*, 2007, pp. 1713–1716.

[5] K. Thambiratnam and F. Seide, "Fragmented context-dependent syllable acoustic models," in *Proc. Interspeech*, 2008, pp. 2418–2421.

[6] A. Hämäläinen, L. ten Bosch, and L. Boves, "Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider," *Speech Communication*, vol. 51, no. 2, pp. 130–150, February 2009.

[7] R. W. M. Ng and K. Hirose, "Syllable: A self-contained unit to model pronunciation variation," in *Proc. ICASSP*, 2012.

[8] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proc. ICSLP*, 1996, pp. 1261–1264.

[9] F. Pellegrino and R. Andre-Obrecht, "Automatic language identification: an alternative approach to phonetic modeling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[10] W. M. Fisher, *Syllabification software*, http://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z, June 1997.

[11] J.-L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1904–1911, August 2007.

[12] "TIMIT acoustic-phonetic continuous speech corpus," National Institute of Standards and Technology, Gaithersburg, MD, 1990.

[13] Jon Fiscus, *Spring 2007 (RT-07) Rich Transcription Meeting Recognition Evaluation Plan*, U.S. NIST, 2007.

[14] T. Hain et al., "The AMIDA 2009 meeting transcription system," in *Proc. Interspeech*, 2010.