ROBUST SPEAKING RATE ESTIMATION USING BROAD PHONETIC CLASS RECOGNITION

Jiahong Yuan and Mark Liberman

University of Pennsylvania

ABSTRACT

Robust speaking rate estimation can be useful in automatic speech recognition and speaker identification, and accurate, automatic measures of speaking rate are also relevant for research in linguistics, psychology, and social sciences. In this study we built a broad phonetic class recognizer for speaking rate estimation. We tested the recognizer on a variety of data sets, including laboratory speech, telephone conversations, foreign accented speech, and speech in different languages, and we found that the recognizer's estimates are robust under these sources of variation. We also found that the acoustic models of the broad phonetic classes are more robust than those of the monophones for syllable detection.

Index Terms— Speaking rate estimation, syllable detection, robustness, broad phonetic class

1. INTRODUCTION

Robust speaking rate estimation can be useful in automatic speech recognition and speaker identification [1], and accurate, automatic measures of speaking rate are also relevant for research in linguistics, psychology, and social sciences. Speaking rate has been found to be related to many factors: individual, demographic, cultural, linguistic, psychological and physiological [2]. For example, human perception experiments using resynthesised stimuli have suggested that speaking rate, but not fundamental frequency, is a perceptually relevant cue to age in voice [3]. Robust speaking rate estimation is crucial for understanding the effects of these factors.

One approach for building a speaking rate estimator that is robust to speaker, genre, dialect and language is to utilize the algorithms developed in syllable detection studies, which were mainly based on energy and periodicity measurements. In an early, influential study, Mermelstein (1975) used a convex hull algorithm on energy between 500 Hz and 4k Hz to locate syllable boundaries [4]. In more recent studies of this kind, Xie and Niyogi (2006) applied a modified convex hull algorithm on both periodicity and normalized full-band energy, one after another, for syllable nuclei detection [5]. Howitt (2000) incorporated Neural Network into a vowel landmark detector using Mermelstein's convex-hull algorithm [6]. His study also demonstrated that energy in a fixed frequency band (300 to 900 Hz) was as good for finding vowel landmarks as the energy at the first formant. Morgan and Fosler-Lussier (1998) used both the first spectral moment of full-band energy and compressed sub-band energy correlation in their algorithm of syllable detection [7]. Wang and Narayanan (2007) extended the sub-band correlation by including temporal correlation and the use of prominent spectral sub-bands for syllable detection algorithm [8]. Zhang and Glass (2009) applied sinusoid fitting on energy peaks to predict possible regions where syllable nuclei can appear, and then a simple slope based peak counting algorithm was used to get the positions of the syllable nuclei [9].

Although these syllable detection algorithms work well on short and fluent speech, e.g., TIMIT, it remains a challenge for the algorithms to handle disfluencies, long pauses, and non-speech segments such as noise and laughter contained in conversational and long speech, e.g. Switchboard. To test their algorithms on Switchboard, for example, [7] and [8] segmented the utterances into short spurt regions based on the manually labelled pause and noise markings, instead of using the entire utterances.

Another approach for speaking rate estimation would be through the use of automatic speech recognition. However, the performance of ASR is much affected by speaking rate, and it has been argued that independent speaking rate estimation is needed for speech recognition [1, 8]. Furthermore, although ASR works well when the training and test data are from the same speech genre, dialect, or language, it's impractical to find or build a recognizer using data that match the test conditions for every task of speaking rate estimation, especially for less common speech genres or languages.

For speaking rate estimation, however, what is important is not the recognition word error rate (WER) or phone error rate. A recognizer that can distinguish between vowels and consonants, e.g., a broad phonetic class recognizer, or a vowel detection algorithm [10, 11], would be sufficient for estimating speaking rate. The broad phonetic classes, e.g., nasals, stops, vowels, etc., possess more distinct spectral characteristics than the phones within the same broad phonetic classes. In a phoneme recognition study [12], it was found that almost 80% of misclassified phonemes were within the same broad phonetic class: vowels/semi-vowels, nasals/flaps, stops, weak fricatives, strong fricatives, and closures/silence. Broad phonetic classes have been applied for improved phone recognition [13, 14], and have been shown to be more robust in noise [14, 15]. Broad phonetic classes have also been used in large vocabulary ASR to overcome the issue of data sparsity and robustness. In the standard triphone acoustic model building process, for example, the last step is to cluster and tie states in order to share data and reduce the number of parameters to be estimated. This is usually done through decision tree-based clustering with broad phonetic classes [16].

In this paper we present experiments of using a broad phonetic class recognizer for syllable detection and speaking rate estimation. The experiments demonstrated the robustness of broad phonetic class recognition for estimating speaking rate. In Section 2, we describe the data and the procedure used for building the recognizer, and in Section 3 and 4 we present the performance of the recognizer on a variety of data sets. Finally, Section 5 contains some concluding remarks.

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

2. A BROAD PHONETIC CLASS RECOGNIZER

A broad phonetic class recognizer was built using the SCOTUS corpus, the CMU pronouncing dictionary, and the HTK Toolkit. The SCOTUS corpus includes more than 50 years of oral arguments from the Supreme Court of the United States. Seventy-eight hour-long arguments from the 2001 term were transcribed and manually word-aligned. We extracted 34,656 speaker turns from these arguments and used them to train the models. The speaker turns were first forced aligned using the Penn Phonetics Lab Forced Aligner [17], and then, the aligned phones were mapped to broad phonetic classes before training. The mappings between the CMU dictionary phone set and the broad phonetic classes are listed in Table I.

TABLE I. BROAD PHONETIC CLASSES AND FREQUENCIES IN TRAINING DATA

Class	Phonetic categorization	CMU phones	Number of tokens
V1	Stressed vowels	Vowel classes: 1 and 2	447,665
V0	Non-stressed vowels	Vowel class: 0	336,278
S	Stops and affricates	B CH D G JH K P T	418,994
F	Fricatives	DH F HH S SH TH V Z ZH	352,968
N	Nasals	M N NG	208,178
G	Glides and liquids	L R W Y	203,683
P	Pauses and non-speech	--	149,268

The acoustic models are GMM-based, mono broad-class three-state HMMs. Each HMM state has 64 Gaussian Mixture components on 39 PLP coefficients (12 Cepstral coefficients plus energy, and Delta and Acceleration). We trained two sets of acoustic models, one for broad-band speech and the other for narrow-band speech, by downsampling the original 44.1 kHz waveforms to 16k Hz and 8k Hz respectively. We also trained a simple "language" model, i.e., broad-class bigram probabilities, using the broad-class transcriptions of the training data. The training was done using the HTK toolkit, and the HVite tool in HTK was used for testing.

To use the broad-class recognizer for syllable detection, we simply count the number of vowels, including both V1 and V0, in the recognition output. We tested the recognizer on a variety of speech data, and found that the grammar scale factor for HVite, which post-multiplies the language model likelihoods from the broad-class lattices, was about 2.5 to have the optimal results on syllable detection. In the following experiments, we set the grammar scale factor to be 2.5, and the other parameters of HVite to be their default values.

3. PERFORMANCE ON TIMIT

Many studies of syllable detection have utilized TIMIT. TIMIT does not contain syllable information. The previous studies were not consistent on whether phones such as /en/, /l/, or /axr/ should be considered as syllable nuclei or not. And, there is no standard scoring toolkit for syllable detection evaluation. Therefore, a correct detection in one study may be determined incorrect in another.

We adopt the evaluation method in [5], which is clearly stated and straightforward. Following the method in [5], we first find the middle points of the V1 and V0 segments from the recognition output. Then, a point is counted as correct if it is located within a syllabic segment (i.e., all vowels plus /el/, /em/, /en/, and /eng/), otherwise, it is counted as incorrect. If two or more points are located within a syllabic segment, only one of them is counted as correct and the others as incorrect. The incorrect points are insertion errors, and the syllabic segments that don't have any correct points are deletion errors. Deletion and insertion error rates are both calculated against the number of syllabic segments in the testing data. There are 1,344 utterances and 17,190 syllabic segments in the testing data, which includes all the utterances in TIMIT test dataset excluding SA1 and SA2 utterances.

We note that this scoring method over-estimates deletion errors, because TIMIT transcribes several common /r/- and /l/-final syllables with sequences like "wire" as [w ay axr].

The results, compared with [5], are shown in Table II. Table II also shows the results from a general monophone recognizer that was built using the same training data as used for the broad class model. It includes 69 monophones, and the monophone models are GMM-based, three-state HMMs. Each HMM state has 32 Gaussian Mixture components on 39 PLP coefficients. A "language" model, i.e., monophone bigram probabilities, was also trained using the same training data. The grammar scale factor used in the monophone recognition was 3.5, which generated the best result.

TABLE II. PERFORMANCE ON TIMIT

	Del. Error	Ins. Error	Total Error
Broad class (with "language" model)	16.0%	8.0%	**24.0%**
Broad class (acoustics only)	14.4%	13.4%	**27.8%**
Xie & Niyogi 2006 ([5])	18.4%	10.9%	**29.3%**
Monophone (with "language" model)	13.0%	9.4%	**22.4%**
Monophone (acoustics only)	7.9%	29.7%	**37.6%**

From Table II we can see that our approach of using broad phonetic class recognition for syllable detection (24% total error) can achieve more than 5% absolute error reduction, about 18% relative reduction, compared to the algorithm in [5] (29.3% total error), which applies a peak detection on both energy and periodicity. When only acoustics is used and no language model is involved, the performance of the broad phonetic class recognizer is slightly better than [5] (27.8% vs. 29.3%).

We can also see from Table II that although monophone recognition is slightly better than broad phonetic class recognition for syllable detection when the "language" model scales are tuned to the optimal values (22.4% vs. 24.0%), it is much worse when no language models are involved (37.6% vs. 27.8%). This result shows that the acoustic models of the broad phonetic classes are more robust than those of the monophones for syllable detection.

Table III summarizes the syllable detection errors from using the broad phonetic class recognizer with no language models involved. The entire TIMIT dataset, 6300 utterances and 241,225 segments (including 80,856 syllabic segments), was used for the error report. The errors can be grouped into three types: “deletions”, i.e., the "gold standard" has the vowel V and the algorithm missed it; “outside insertions”, i.e., the algorithm found a vowel whose midpoint is not inside any vowel segment in the gold standard; and “inside insertions”, i.e., the algorithm found two or more vowels whose midpoints are inside the same vowel segment in the gold standard.

There were totally 7,448 outside insertions. About half of the outside insertions (3635, 48.8%) occurred at a glide (/y, w/) or a liquid (/r/, l/). Besides, 1411 (18.9%) of the outside insertions occurred at /q/, a glottal stop that “may be an allophone of t, or may mark an initial vowel or a vowel-vowel boundary” (TIMIT documentation).

The deletion and inside insertion errors are summarized in Table V. We can see that, as expected, the syllabic nasals and laterals, /el, em, en, eng/, and the schwa vowels, /ax, ax-h, ax-r/, are more likely to be deleted; and the diphthongs, /aw, ay, ey, ow, oy/, are more likely to have inside insertions. It is not clear, though, why /ao/ is more likely to be deleted than the other vowels whereas /ux/, which is fronted /uw/, is more likely to have inside insertions.

TABLE III. DELETIONS AND INSIDE INSERTIONS

	Total	Deletions	Inside insertions
aa	4197	422 (0.10)	178 (0.04)
ae	5404	146 (0.03)	437 (0.08)
ah	3185	323 (0.10)	107 (0.03)
ao	4096	***1107 (0.27)***	164 (0.04)
aw	945	12 (0.01)	82 (0.09)
ax	4956	996 (0.20)	14 (0.00)
ax-h	493	277 (0.56)	1 (0.00)
axr	4790	1599 (0.33)	161 (0.03)
ay	3242	110 (0.03)	347 (0.11)
eh	5293	570 (0.11)	203 (0.04)
el	1294	388 (0.30)	24 (0.02)
em	171	116 (0.68)	1 (0.01)
en	974	525 (0.54)	10 (0.01)
eng	43	24 (0.56)	1 (0.02)
er	2846	872 (0.31)	294 (0.10)
ey	3088	113 (0.04)	253 (0.08)
ih	6760	857 (0.13)	197 (0.03)
ix	11587	1988 (0.17)	111 (0.01)
iy	9663	915 (0.09)	515 (0.05)
ow	2913	277 (0.10)	343 (0.12)
oy	947	107 (0.11)	347 (0.37)
uh	756	98 (0.13)	31 (0.04)
uw	725	113 (0.16)	74 (0.10)
ux	2488	218 (0.09)	***512 (0.21)***

4. ROBUSTNESS TO SPEEN GENRE AND LANGUAGE

To evaluate the robustness of our method, we used the same broad phonetic recognizer for estimating speaking rate in English telephone conversations, foreign accented English speech, and Mandarin Chinese speech.

Both [7] and [8] tested their algorithms for speaking rate estimation using the ICSI manual transcription portion of the Switchboard telephone conversation speech. The transcriptions contain syllabic boundary markings. In [7] and [8], the utterances were segmented into spurt regions using the pause and noise markings in the transcriptions, and the spurts were used for both training and testing. In our experiment, we did not cut utterances into spurts. Instead, we ran the broad class recognizer on the entire utterances, and let the recognizer handle pauses and non-speech segments in the utterances. We tested on the WS-97 release of the ICSI Switchboard data. It has 5119 utterances in total. To calculate the detected speaking rate, we simply counted the number of vowels, both V1 and V0, in the recognition of an utterance, and divided the number by the length of the utterance. This detected syllable rate was compared with the reference rate, i.e., the number of transcribed syllables divided by the length of the utterance. Following [7] and [8], the correlation between the two rates was used for evaluation, as well as the mean error and the standard deviation of the errors. The results, compared with [7] and [8], are shown in Table IV.

TABLE IV. PERFORMANCE ON SWITCHBOARD

	Correlation	Mean Error	Stddev Error
Broad class	**.763**	**-.161**	**0.780**
Wang & Narayanan 2007 ([8])	.745	.339	0.796
Morgan & Fosler-Lussier 1998 ([7])	.671	.464	1.121

Clearly, the performance of the broad class recognizer is comparable to the state-of-the-art algorithm in [8]. In [8], the algorithm learns the optimal settings of many parameters from data similar to the test data. Our approach, however, does not need to tune any parameters or retrain the acoustic or language model. Furthermore, unlike the previous algorithms, the broad class phonetic recognizer can automatically handle pauses and non-speech segments. This presents a great advantage for estimating speaking rate in natural speech.

To test on foreign accented speech, we utilized the CSLU Foreign Accented English corpus, which includes English spoken by native speakers of 22 languages, who talked about themselves in English for up to 20 seconds. Three native speakers of American English independently listened to each utterance and judged the speakers' accents on a 4-point scale, from negligible/no accent (1), to very strong accent (4). We randomly selected 200 recordings from the corpus, including eight L1 languages: Cantonese, French, German, Japanese, Mandarin, Russian, Spanish, and Vietnamese, and a wide range of accent levels. The CSLU corpus does not provide word transcriptions. We manually transcribed the 200 recordings, and calculated a reference rate for each recording based on the transcription. We then applied the broad phonetic class recognizer on the recordings, and calculated the detected rate by dividing the number of detected vowels by the duration of the recording. The results are as follows: the correlation is 0.898; the mean error is -0.01; and the standard deviation of the errors is 0.36.

Figure 1 illustrates the results for different L1 languages and accent levels. We can see that the speaking rate estimation is robust to foreign accented speech, it is not severely affected by either L1 or accent level.

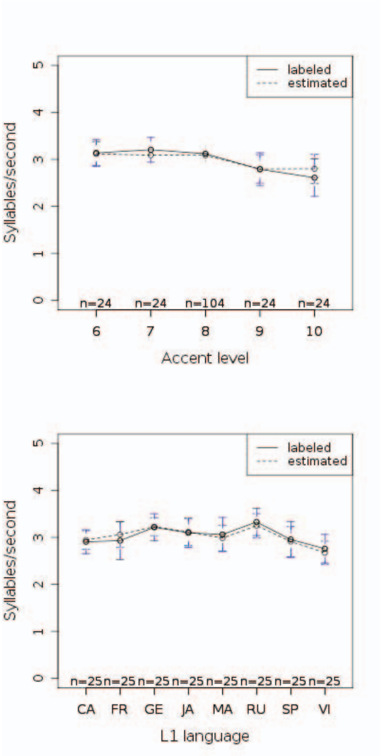

Fig. 1. Performance on different accent levels (top) and L1 languages (bottom). Accent level is the sum of three judges on a 4-point scale.

Because the language model has only mild effect on syllable detection when using the broad class recognizer, we predict that the recognizer can also be used for speaking rate estimation in other languages. Although different languages have different syllable structures, and therefore, different broad-class bigram probabilities, the broad phonetic class recognizer can be used without the language model if no such model is available for the test language, or if the test language is unidentified.

Figure 2 presents the results of speaking rate estimation using the broad phonetic class recognizer on Mandarin Chinese. The test data were randomly selected, 5,000 utterances from the Hub-4 Mandarin Broadcast News corpus. No language models were involved in the test.

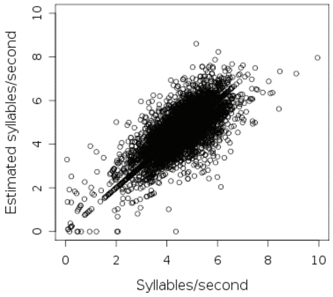

Fig. 2. Performance on Hub-4 Mandarin Broadcast News: Correlation: **.755**; Mean Errors **.055**; Stddev Error: **.730**.

5. CONCLUSION

We built a broad phonetic class recognizer, and applied it for syllable detection and speaking rate estimation. Its performance is comparable to state-of-the-art syllable detection and speaking rate estimation algorithms, and it is robust to different speech genres and different languages. Our broad class acoustic models are more robust than monophone models for syllable detection. With no language models involved, the broad class recognizer still has good performance on syllable detection and speaking rate estimation, which opens up many application opportunities.

6. ACKNOWLEDGEMENTS

We thank Rebecca Marsh for transcribing the foreign accented speech data. This work was partially supported by NSF award 0325739.

7. REFERENCES

[1] N. Morga, E. Fosler, and N. Mirghafori, "Speech Recognition Using On-Line Estimation Of Speaking Rate," *Eurospeech* 1997, pp. 2079-2082.

[2] J. Yuan, M. Liberman, and C. Cieri, "Towards an Integrated Understanding of Speaking Rate in Conversation," *Interspeech 2006*, pp. 541-544.

[3] J. Harnsberger, R. Shrivastav, W. Brown Jr., H. Rothman, and H. Hollien, "Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age," *Journal of Voice*, 22 (1), 2008, pp. 58-69.

[4] P. Mermelstein. "Automatic segmentation of speech into syllabic units," JASA, 58(4), 1975, pp. 880-883.

[5] Z. Xie and P. Niyogi, "Robust Acoustic-based Syllable Detection," *Interspeech 2006*.

[6] A. W. Howitt, *Automatic syllable detection for vowel landmarks*, PhD thesis, MIT, 2000.

[7] N. Morgan and E. Fosler-Lussier, "Combining Multiple Estimations of Speaking Rate," *ICASSP 1998*, pp. 729-732.

[8] D. Wang and S. Narayanan, "Robust Speech Rate Estimation for Spontaneous Speech," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 15, 2007, pp. 2190-2201.

[9] Y. Zhang and J. Glass, "Speech Rhythm Guided Syllable Nuclei Detection," I*CASSP 2009*.

[10] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," ICASSP 1998, pp. 945-948.

[11] F. Pellegrino, J. Farinas, and jl. Rouas, "Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech," Speech Prosody 2004, pp. 517-520.

[12] A. Halberstadt and J. Glass, "Heterogeneous Acoustic Measurements for Phonetic Classification," *Eurospeech, 1997*, pp. 401-404.

[13] O. Scanlon, D. Ellis, and B. Richard, "Using Broad Phonetic Group Experts for Improved Speech Recognition," *IEEE Tran. on Audio, Speech, and Language Processing*, Vol. 15, 2007, pp. 803-812.

[14] T. N. Sainath and V. Zue, "A comparison of Broad Phonetic and Acoustic Units for Noise Robust Segment-Based Phonetic Recognition," *Interspeech 2008*, pp. 2378-2381.

[15] T. N. Sainath, D. Kanevsky, and B. Ramabhadran, "Broad Phonetic Class Recognition in a Hidden Markov Model Framework using EBW Transformations," *ASRU 2007*, pp. 306–311.

[16] S. Young, J. Odell, and P. Woodland, "Tree-based State Tying for High Accuracy Acoustic Modelling," *Proc. ARPA Human Language Technology Conf.* 1994.

[17] J. Yuan and M. Liberman, "Speaker Identification on the SCOTUS Corpus," *Acoustics 08*, pp. 5687-5690.