Glide landmark detection using band-limited energy ratio contours

Soojin Park, Jeungyoon Choi, Honggoo Kang
(*Dept. of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea*)

Abstract: A detection system for American English glides /w y r l/ in a knowledge-based automatic speech recognition system is presented. The method uses detection of dips in band-limited energy to total energy ratios, instead of detecting dips along the unmodified band-limited energy contours. By using band-limited energy ratio, the dip detection is applicable in not only intervocalic regions but also in non-intervocalic regions. A Gaussian mixture model(GMM) based classifier is then used to separate the detected vowels and nasals. This approach is tested using the TIMIT corpus and results in an overall detection rate of 69.5%, which is a 4.7% absolute increase in detection rate compared with an hidden Markov model (HMM) based phone recognizer.

Key words: landmarks; glide detection; knowledge-based speech recognition

CLD number: TN912.34 **Document code**: A
Article ID: 1674-8042(2012)04-0352-05 **doi**: 10.3969/j.issn.1674-8042.2012.04.011

Current automatic speech recognition (ASR) systems typically use statistical approaches, i.e. hidden Markov models (HMMs) to capture patterns in the speech signals. These models usually assume that the speech frames are independent so that each frame can be analyzed. However, phonetic information of a speech signal is not uniformly distributed across the whole utterance. Instead, valuable information is found in the vicinity of abrupt spectral discontinuities or transitions. Thus, listeners need not only listen to each of the time segment of a speech carefully but also focus on the instances where more information is located. Various perceptual experiments support the hypothesis that human speech recognition is based on the regions of abrupt change[1-3].

The accumulation of knowledge from linguistics has led to investigations of direct modeling of linguistic knowledge. For example, Stevens[4] outlined a distinctive feature-based speech recognition system that extracted linguistic descriptions of speech sounds from the signal. In this approach, landmarks[5] are first extracted, which are locations in the signals that indicate articulatory configurations or movements involved in the production of broad classes of speech sounds. Distinctive features, which are linguistic descriptions of speech sounds[6], are then extracted from around the landmarks. The landmark-based ASR system has several advantages over statistical approaches. Since detailed analysis is carried out only in specific regions designated by the landmarks, the system is supposed to be more efficient. Besides that, different resolutions according to the landmarks are applicable. For example, a short temporal window can be applied for a transient burst of a stop consonant while a longer one for a vowel.

Among the landmarks, detection of glides, especially non-intervocalic glides, is the most difficult task. Glide detection has been investigated by Espy-Wilson[7-8], in which glides in a carrier phrase "(word)_pa" are used for training, and dip/peak detection[9] along band energy and formant frequency contours is used for classification. Classification rates are around 79%, when tested on continuous sentences for 15 speakers. Sun[10] detected prevocalic /w/ and /y/ using signal amplitude and first formant energy, resulting in a detection rate of 93.3%, with 6.6% false alarm for isolated intervocalic segments from 3 speakers, and 88.0%, and 9.4%, for continuous sentence utterances from 5 speakers, respectively. However, the methods described in these studies are not directly applicable to general distinctive feature-based systems, and evaluation is carried out on utterances from a limited number of speakers.

This paper investigates an improved method for detection of the glides /w y r l/ for a knowledge-based speech recognition system. Especially, this

* **Received data**: 2012-08-05
Foundation item: The Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by the National IT Industry Promotion Agency (NIPA-2012-H0301-12-2006)
Corresponding author: Soojin Park (weekny@dsp.yonsei.ac.kr)

study improves upon Espy-Wilson's method by proposing dip detection along band-limited energy to total energy contour ratio, to improve detection of non-intervocalic glides.

1 Methods

1.1 Database

The TIMIT database[11] contains 16 kHz sampled recordings of 630 speakers, each reading ten sentences in a noise-free environment. Time-aligned orthographic, phonetic and word transcriptions for each utterance are provided. According to the sonority hierarchy[12], glides appear next to vowels. Therefore, locations of vowels may be used as anchors to detect glides.

In this study, we assume vowel landmarks have been found in advance, and our goal is to detect glides in all sonorant regions that abut vowel landmarks. This assumption is plausible since vowel landmarks can be detected with high accuracy[13]. Although it is expected that there will be errors in detecting such landmarks in a practical system, in order to evaluate glide detection performance in the absence of endpoint detection errors, phone labels provided in TIMIT are used as references in this study. That is, all sonorant regions, including all vowels, semivowels, liquid and nasals, are first identified by using TIMIT phone labels.

First, it is confirmed whether glides appear adjacent to vowels in the database. Examination of the TRAIN set portion of the TIMIT corpus showed 3 /w/s, 19 /r/s, and 6 /l/s which are not adjacent to vowels. However, a closer look shows these are followed by the syllabic nasals /em/ or /en/, which may constitute syllabic nuclei, and may be considered vocalic segments. A similar distribution is observed for the TEST portion as well. To simplify modeling, these cases are treated as exceptions and are not included in this study.

The remaining sonorant regions may be categorized into three different groups according to the position relative to vowels: intervocalic, prevocalic and postvocalic. Each region may contain one or more glides, or none (e.g. intervocalic nasals or two concatenated vowels). Counts of sonorants within each type of region TIMIT are given in Table 1.

Table 1 Counts of glides within intervocalic, prevocalic and postvocalic sonorant regions for training and test sets of the TIMIT database. Glides are denoted w, y, r and l

	intervocalic				prevocalic				postvocalic			
	w	y	r	l	w	y	r	l	w	y	r	l
Train	1 405	874	2 1342	802	1 739	841	2 8531	910	1	0	1 5331	083
Test	591	320	890	1105	646	313	1040	763	0	0	593	483

The proposed glide detection system consists of two steps: dip detection along band-limited energy contours and separating out vowels and nasals using classifiers. Each step is described below.

1.2 Dip detection

English semivowels /w y r l/ have low first formant (F_1) due to oral cavity narrowing. The decrease in F_1 involves a reduction in the spectral amplitude not only in the amplitude of F_1 peak in the spectrum but also in those of the higher-frequency spectral peaks[14]. This phenomenon can be observed as a dip in low to mid-frequency band limited energy has been used to detect glides. Band frequencies are selected as 640 − 2 800 Hz (E_1) and 2 000 − 3 000 Hz (E_2)[7-8]. The E_1 band is chosen to include the second and the third formants (F_2 and F_3), which have weaker amplitudes compared with its adjacency to vowels in the case of glides. The E_2 band is chosen to aid the detection of /r/, which does not form notable dips in the 640 − 2 800 Hz frequency range. The band energies are calculated with a 25-ms window at 10-ms intervals.

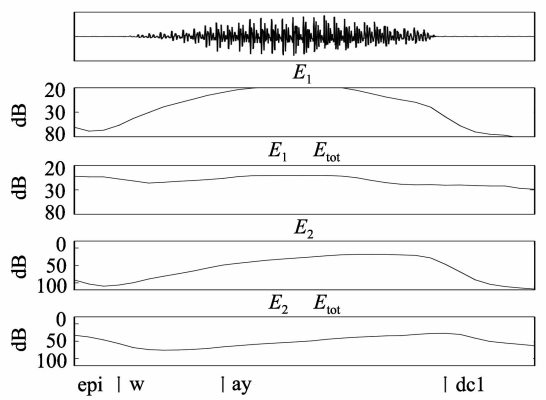

Fig. 1 An example of energy dip formed in the pronunciation of "wide" following an epenthetic silence. E_1, E_2 and E_{tot} refer to band-limited energies in dB in 640 − 2 800 Hz, 2 000 − 3 000 Hz and the total band, respectively

For intervocalic regions, dip detection in band energy is highly reliable since neighboring vowels form energy peaks. For prevocalic and postvocalic regions, glides are often placed adjacent to silence or stop closures and are hidden in the overall energy reduction. Therefore, using Espy-Wilson's definition of dip depth as energy difference between the minimum energy in consonants and that of an adjacent vowel becomes less effective. Thus, in this paper, we propose subtracting the total log-energy of a given frame from the band limited log-energy to compensate for the reduction in the total energy. That is, band-limited energy to total energy ratio is used instead of the band-limited energy itself.

Fig. 1 shows a sample from the TIMIT data set. The word "wide" is pronounced following an epenthetic silence. Energy dips at /w/ are not observed in the E_1 and E_2 contours but are presented in the E_1-E_{tot} and E_2-E_{tot} contours.

1.3 Separation of vowels and nasals

Once candidate locations for glides are found using dip information, the locations are examined for the presence of glides and vowels and nasals are separated out. The dip depth information and band-limited energies are used as features. 0 − 900 Hz band (E_0) is included as a low frequency band since vowels are expected to form peaks in the low frequency band[15].

Along with band energies, the additional measurements related to the vocal tract are included. Widely used measurements such as the first three formant frequencies (f_1, f_2 and f_3), amplitudes (A_1, A_2 and A_3) and bandwidths (B_1, B_2 and B_3) are found. Although vocal tract measurements have been mostly used in speech recognition, the glottal source has substantial interactions with the vocal tract[14]. Therefore, the features related to voice source information such as open quotient, spectral tilt, harmonic amplitudes and harmonics-to-noise ratios are found. Open quotient refers to the ratio of the interval in which the vocal folds are open to the total pitch period, and spectral tilt is defined as the slope of a least squares linear fit to the log power spectrum. Hanson[16] suggested that open quotient could be estimated by calculating H_1-H_2, and spectral tilt by calculating H_1-A_3. In this study, the measures H_1^*-H_2^* and H_2^*-H_4^* are found for open quotient, and spectral tilt measures include H_1^*-A_1^*, H_1^*-A_2^* and H_1^*-A_3^*. Asterisks indicated that spectral magnitudes had been corrected for formant effects[17]. Finally, harmonics-to-noise ratios (HNRs) from 4 frequency bands are chosen. They are: 0 − 500 Hz, 0 − 1 500 Hz, 0 − 2 500 Hz and 0 − 3 500 Hz.

All formant frequency, amplitude and bandwidth values, as well as voice source features, are normalized by their means and standard deviations for each utterance. In this study, values for fundamental frequency, formant frequencies, amplitudes and bandwidths are extracted using the VoiceSauce toolkit[18] with frame length 25 ms and step size 10 ms. The features are then culled from an ANOVA test to select significant measurements with $p < 10^{-5}$.

Using these measurements, Gaussian mixture models (GMMs) are used for classification. Eight mixtures are sufficient to model distributions. Although GMMs may be used directly for multi-class classification, in this study, they are used in tree-based cascading two-class classifiers, in order to alleviate the data imbalance problem. The first step separates vowels from glides, and the second separates nasals from glides.

2 Experimental results

2.1 System performance

In this paper, glide landmark detection performance is evaluated with a 10-ms tolerance from the reference TIMIT phone labels. As expected, glides in intervocalic regions are the most reliably detected, with a detection rate of 71.0%. Prevocalic glides show similar performance of 69.6%, while postvocalic glides show the lowest rates of 65.2%. Overall system performance is summarized in the following Table 2.

Table 2 Glide detection results. The last three columns indicate classification results, with rates in parentheses. Insertion denotes occurrences of two or more detections within a single glide

	No. tokens	Detected	Glide	Vowel	Nasal
			Intervocalic		
Glide	2 906	2 694(92.7%)	2 063(76.6%)	407(15.1%)	224(8.3%)
Insertion		1 338	508(38.0%)	573(42.8%)	257(19.2%)
Vowel		5 452	1 272(23.3%)	4 034(74.0%)	146(2.7%)
Nasal	1 910	2 269	338(14.9%)	540(23.8%)	1 391(61.3%)
			Prevocalic		
Glide	2 762	2 597(94.0%)	1 921(74.0%)	456(17.6%)	220(8.5%)
Insertion		566	198(35.0%)	306(54.1%)	62(1.0%)
Vowel		10 275	1 571(15.3%)	8 350(81.3%)	354(3.4%)
Nasal	717	871	82(9.4%)	153(17.6%)	636(73.0%)
			Postvocalic		
Glide	1 076	830(77.1%)	701(84.5%)	108(13.0%)	21(2.5%)
Insertion		238	141(59.2%)	80(33.6%)	17(7.1%)
Vowel		8 898	1 850(20.8%)	6721(75.5%)	327(3.7%)
Nasal	2 272	1 674	88(5.3%)	1 119(66.8%)	467(27.9%)

To compare performance, an mel-frequency cepstrum coefficient(MFCC) based 3-state HMM phone recognizer with 8-Gaussian mixture is implemented with the HTK toolkit[19]. The overall detection rate of 69.5% shows improvement over the HMM phone recognizer, which yields a glide detection rate of 64.8%.

2.2 Detection errors

Since a deletion error in the dip detection step is not recoverable, this type of error is crucial for performance. The total deletion rate in the detection step is 9.2%. More specifically, the rates is 7.3%, 6.0% and 22.9% for intervocalic prevocalic, and postvocalic region, respectively.

For intervocalic glides, the formation of the dips is most affected by adjacent glides or nasals.

In this case, a dip that occurs for the adjacent segment often overshadows the dip in the glide. This case comprises about 75% of deleted intervocalic glides. Especially for adjacent nasals, a dip tends to be detected in the nasal instead of the glide and this case comprises about 35% of all intervocalic glide deletions. This problem s not as significant for prevocalic and postvocalic glides, where about 4% of glides adjacent to nasals or other glides are deleted. The most significant factor causing deletion in prevocalic glides is short of duration, which may not allow sufficient time to form a dip during transition. For example, about 52% of deleted prevocalic glides are shorter than 30 ms. This is not as prevalent for intervocalic and postvocalic glides, where about 7% of all deletions are attributed to this effect.

For postvocalic glides, adjacent syllabic liquids /el er axr/ show similar effects as adjacent glides and nasals.

In this case, for about 22% of all the deleted postvocalic glides, an energy dip tends to be placed in the syllabic liquid, resulting in a deletion of the glide as well as an insertion of a vowel. Also, intervocalic glides in such cases showed similar tendencies, contributing to about 9% of deletion errors. More than half of deletions for postvocalic glides (about 58%) are from a glide preceding a stop closure. In this case, the liquids (liquids are the only allowed glides in this case) do not form enough of a fall in band-limited energy contour so that its energy is similar to the preceding vowels. One possibility for overcoming deletion errors may be to employ additional frequency bands. It is observed that when glides are deleted in the contexts described above, other bands such as 1 000 − 2 000 Hz and 1 500 − 2 500 Hz formed dips, which may be used in addition to those used in this study.

2.3 Classification errors

In the classification step, the balanced error rates are 33.7%, 26.2% and 39.2% for intervocalic, prevocalic and postvocalic classifiers, respectively. The classification error rate may be reduced slightly by selecting an optimal number of mixtures for the GMMs. But more fundamentally, it may be necessary to add features tailored to discriminating between glides/vowels and glides/nasals. Also, it is revealed that syllabic liquids contribute a large portion to vowel insertion, with about 46%, 40% and 49% of vowel insertions in intervocalic, prevocalic and postvocalic regions. Therefore, further measurements for filtering out syllabic liquids may also be needed.

Overall, this system has a drawback in its high insertion rate. One possible strategy may be to apply a threshold in the detection step to exclude minor band energy fluctuation from vowels. Preliminary results show that when thresholds are set to preserve about 70% of glides for each contour, about 34% of insertions are removed while retaining 97% of glides. Secondly, posterior probabilities can be calculated for each glide /w y r l/, and observations with no evidence related to glides can be abandoned. In addition, inserted glides may be also be removed at a higher level of processing, such as in the lexical matcher.

3 Conclusion

Glide landmark detection has been a challenging problem, especially for non-intervocalic tokens. In this study, we propose an improved glide landmark detection system that detects dips along band-limited energy to total energy ratio contours. This method can be applied to detection of prevocalic and postvocalic glides as well as intervocalic glides, and the performance shows a notable increase in detection rate compared to an HMM phone recognizer. The insertion rate remains relatively high, but several strategies may be used to alleviate the problem. Further study will include glide landmark detection independent of vowel landmark detection and combination with an HMM based speech recognition system.

References

[1] Jenkins J J, Strange W, Edman T R. Identification of vowels in vowelless syllables. Perception & Psychophysics, 1983, 34(5): 441-450.

[2] Furui S. On the role of spectral transition for speech perception. Jounal of Acoustic Society of America, 1986, 80(4): 1016-1025.

[3] Stevens K N. Evidence for the role of acoustic boundaries

in the perception of speech sounds. Journal of Acoustical Society of America, 1981, 69(s1): s116.
[4] Stevens K N. Toward a model for lexical access based on acoustic landmarks and distinctive features. Journal of Acoustical Society of America, 2002, 111(4): 1872-1891.
[5] LIU S A. Landmark detection for distinctive feature-based speech recognition. Journal of Acoustical Society of America, 1996, 100(5): 3417-3430.
[6] Chomsky N, Halle M. The sound pattern of English, Cambridge MA: MIT Press, 1968.
[7] Espy-Wilson C Y. An acoustic-phonetic approach to speech recognition: Application to the semivowels. Massachusetts Institute of Technology, Ph. D. thesis, 1987.
[8] Espy-Wilson C Y. A feature-based semivowel recognition system. Journal of Acoustical Society of America, 1994, 96(1): 65-72.
[9] Mermelstein P. Automatic segmentation of speech inty syllabic units. Journal of Acoustical Society of America, 1975,58(4): 880-883.
[10] Sun W. Analysis and interpretation of glide characteristics in pursuit of an algorithm for recognition, Massachusetts Institute of Technology, MS Thesis, 1997.
[11] Garofalo J S, Lamel L F, Fisher W M, et al. The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM. Linguistic Data Consortium, 1993.
[12] Hankamer J, Aissen J. The sonority hierarchy. Papers from the parasession on Natural Phonology, Chicago: Chicago Linguistic Society, 1974.
[13] Howitt A W. Vowel landmark detection. Proc. of Interantional Conference on Speech and Language Processing, 2000.
[14] Stevens K N. Acoustic Phonetics, Cambridge MA: MIT Press, 1998.
[15] Espy-Wilson C Y, Pruthi T, Juenja A, et al. Landmark-based approach to speech recognition: An alternative to HMMs. Proc. of Interspeech, Antwerp, Belgium, 2007: 886-889.
[16] Hanson H M. Glottal characteristics of female speakers: acoustic correlates. Journal of Acoustical Society of America, 1997,101(1): 466-481.
[17] Iseli M, Shue Y, Alwan A. Age, sex, and vowel dependencies of acoustic measures related to the voice source. Journal of Acoustical Society of America, 2007, 121(4): 2283-2295.
[18] Shue Y L, Keating P A, Vicenic C. VoiceSauce: a program for voice analysis. Journal of Acoustical Society of Amercial, 2009, 126: 2221.
[19] Young S, Evermann G, Gales M, et al. HTK documentaion. [2012-03-15]. http://htk.eng.cam.ac.uk/.