Acoustics'08 Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Detection of obstruent consonant landmark for knowledge based speech recgonition system

Jung-In Lee and Jeung-Yoon Choi

Yonsei University, 134 Sinchon-dong, Seodaemun-gu, 120-749 Seoul, Republic of Korea
junginida@dsp.yonsei.ac.kr

Obstruent consonant landmarks are detected using spectral energy difference profiles. This study expands upon previous work by Liu. A. [J. Acoust. Soc. Am. 100, 3417-3430]. The proposed algorithm detects four types of landmarks: [stop closure], [stop release], [fricative closure] and [fricative release], where affricates are detected by combining [stop closure], [fricative closure] and [fricative release]. In addition to finding abrupt changes in energy differences, we use energy contours, relative energy and spectral center of gravity differences. This method results in improved performance particularly for CV obstruents. Overall detection rates for stop closure and release are 76.9% and 85.7% for obstruent landmarks in TIMIT, and fricatives yield 82.2% and 83.6% respectively. For strident fricatives, the figures are 94.7% and 93.6%.

1 Introduction

This paper expands upon previous study about landmark detection for knowledge-based speech recognition. An earlier implementation detects the glottis, sonorant, and burst landmarks [1]. Each landmark can be obtained by finding peaks and combining the extreme values in the sub-band energy. A burst landmark indicates a location where abrupt changes occur in the whole sub-band energies. However, it sometimes misses out the landmarks when affected by adjacent phonemes. The proposed method focuses on finding of burst landmarks, and refines the detection method to improve the performance of obstruent consonant landmark detection.

Landmark detection is a first step of a knowledge-based speech recognition system described by Stevens et al [2, 3]. The landmark offers timing information related to the manner, and the later steps are performed in the vicinity of landmarks. A knowledge-based approach is flexibly built by the system designer, thus the appropriate features or parameters for an acoustic analysis are chosen based on the information of acoustics and linguistics. The knowledge-based approach attempts to construct a model of human perception process.

The second step of knowledge-based speech recognition is extraction of distinctive features at the vicinity of the landmarks. Distinctive features denote the minimal set of features that make linguistic discriminations. The features are selected to represent the acoustical characteristics of speech [4, 5]. The decision of phoneme identity is derived from the binary values of features. The later step combines the phonemes represented by features, and then hypothesizes the word by comparing the sequence of phonemes and a lexicon. Theoretically, all phonemes are represented by the set of distinctive features straightforwardly, but the unified system for the extraction of each feature from speech is not completed.

Landmarks are strongly related to the manner of speech, and they are subdivided into four types: abrupt-consonantal, abrupt, non-abrupt, and vocalic. The detection algorithm of first two landmarks is implemented by Liu. A [1]. The landmarks of abrupt-consonantal and abrupt landmarks are associated with glottis vibration, obstruent consonant, and sonorant consonant. The glottis vibration can be regarded as an onset or an offset of voicing. They are assigned as **g** (glottis landmark), **s** (sonorant landmark), and **b** (burst landmark). Main cues of each landmark are to extract the extreme changes in the specific sub-band energy. However, it does not perfectly cover the case where vowel and obstruent consonant occur sequentially. For example, in TIMIT corpus, there is relatively high energy above 4 kHz for both vowel and consonants, thus the abruptness does not always occur~~s~~ in whole sub-band energies at the boundaries of vowels and consonants.

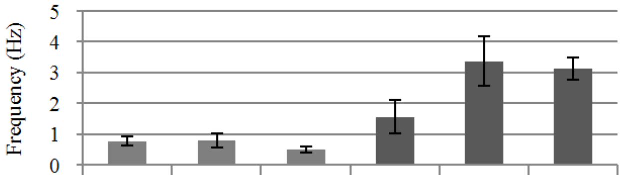
This study focuses on the problems of earlier study and applies additional parameters in order to improve the performance of obstruent consonants. Proposed system is constructed based on the landmark detection proposed by Liu, and additional processing operates in parallel. Decision process used to select the reliable landmarks are included in final step. The obstruent consonant closure, release, and stop closure landmark are extracted from the proposed method. Fricative closure, stop release, affricates closure are associated with the obstruent consonant closure, and fricative release and affricate release are matched to the obstruent consonant release.

The remainder of this paper is organized as follows. Section 2 describes parameters used in the proposed method of landmark detection. The detail about detection method of obstruent consonant landmarks is described in section 3. Experimental results of proposed method are given in section 4. Finally, we summarize and conclude the paper.

2 Acoustic parameters

The parameters used in the proposed refinement method are specified as RMS energy, band RMS energy, spectral center of gravity, and voiced probability. Additional parameters are applied to solve the problems in the previous landmark detection algorithm.

Earlier study of landmark detection uses the peak amplitude of each sub-band as sub-band energy, because the peak amplitude for the range of the formant frequency represents the movement of formant approximately. Otherwise, the RMS energy averages the peaks and nulls in the power spectrum, thus it is not adequate to detect the abrupt change in the sub-band. However, the RMS energy efficiently monitors the changes of energy in perspective. Thus, the RMS energies are added to the proposed system to improve the performance of landmark detection. RMS energy is additionally used to detect the landmark of stop closure, and band RMS energy between 5.0 to 8.0 kHz is used to detect obstruent burst landmarks. Abrupt changes are calculated by the peaks of the extreme values from the first order difference of each RMS energies. Spectral center of gravity and voiced probability are described in detail next.



	vowel	glide	nasal	stop	fricative	affricate
SCOG	784.21	796.52	510.39	1565.22	3372.26	3139.05

Fig.1 Average values of SCOG for sx sentences of training set in TIMIT

2.1 Spectral center of gravity

The center of gravity in the power spectrum determines the dominant frequency where the energy is concentrated [6]. The pattern of spectral center of gravity (SCOG) is similar to the zero-crossing rate (ZCR) or the frequency of maximum peak amplitude (MPF). However, the SCOG is influenced by both maximum amplitude and the distribution of energy in the power spectrum. Therefore, the SCOG is more efficient to monitor the specific changes in the spectrum than ZCR and MPF. SCOG is easily calculated by

$$SCOG(i) = \frac{\sum_{n=1}^{N} nS_i(n)}{\sum_{n=1}^{N} S_i(n)} \qquad (1)$$

where $S_i(n)$ is a power spectrum with frequency index *n* at the frame index *i*, respectively. As shown in Fig. 1, the spectrum of periodic signal such as vowel, glide, and sonorant is weighted to the low frequency bands, but the spectrum of obstruent consonant tends to be weighted above 1.5 kHz. The results of Fig. 1 are preliminarily obtained based on the sx groups of training set in the TIMIT corpus. According to the experimental result of SCOG, the thresholds used in the decision step are determined.

The extreme values in the first order difference of SCOG mean that the center of weight in the power spectrum changes abruptly. Therefore, the difference of SCOG can detect the landmarks which are missed from the sub-band energy profiles.

2.2 Voiced probability

Voiced probability is used to increase the robustness to determine the voicing region by applying a two-state model [7, 8]. Most obstruent consonants are generally classified to the unvoiced signal. The pairs of glottis landmarks also represent the regions of voicing. However, the hard decision based on the glottis landmark causes a cumulative error from the glottis landmark detection. Voiced probability provides reliable decision of voicing independent to the result of glottis landmark detection.

The features used to make Gaussian model consist of normalized cross-correlation values for lag 1, normalized cross-correlation for pitch lag, RMS energy, and relative energy with maximum energy in the signal. In this study, the fixed training model is used to calculate the voiced probability. Voiced probability gets values around 1.0 for the vowels, glides, and liquid. Otherwise, the probability is almost 0 for the obstruent consonants.

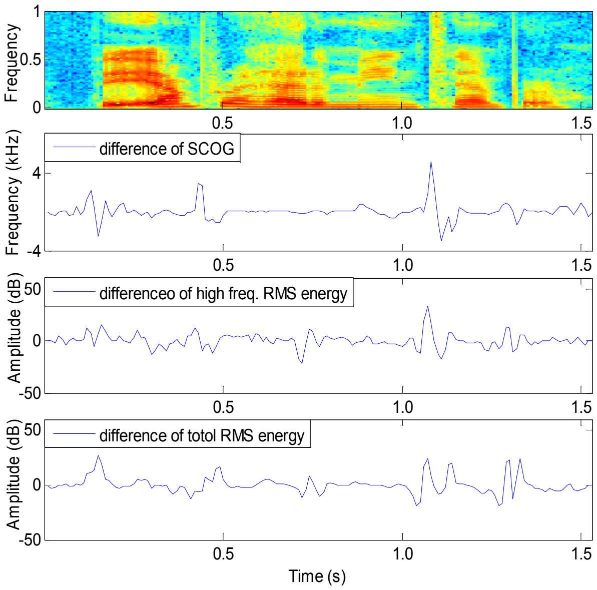


Fig.2 Difference values of additional parameters TIMIT sx127: 'The emperor had a mean temper.'

3 Detection method

3.1 Candidates of landmarks

The peak values of first order difference of additional parameters are combined with the burst landmark extracted by Liu's method as candidates of landmarks. There are burst landmarks, peaks of difference of SCOG, high frequency band RMS energy (5.0-8.0 kHz), and negative peaks of total RMS energy in the candidates. The peaks of difference of SCOG and high frequency band energy are used in the detection of obstruent closure and release landmarks with burst landmarks. Negative peaks of total RMS energy are only used in the detection of stop closure landmarks. Fig. 2 shows an example of the first order of difference of additional parameters. The repetitions of peaks are rejected in the candidates.

The burst landmarks are extracted by combining the peaks locations extracted by picking the extreme values from the whole sub-band energies. It means that the abrupt changes for whole frequency bands are regarded as a burst landmark associated with the obstruent consonant. Positive peaks marked as '**+b**' are related to the locations for fricative closure and stop release, and negative peaks marked as '**-b**' are related to locations of fricative release and stop closure [1].

Candidates of obstruent landmarks are tested with the development set which consists of 100 sentences from the sx groups in the training set of TIMIT corpus. About 90% of landmarks are matched to the transcription of TIMIT, respectively. However, the preliminary test does not take into account insertions. In order to reduce needless landmarks from the candidates, decision rules are required for the landmark detection system.

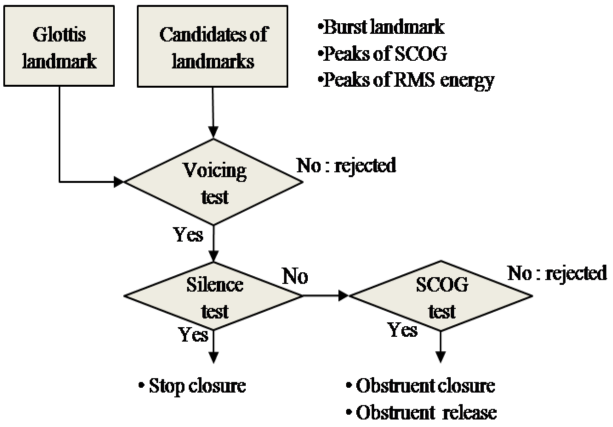


Fig.3 Decision rules of proposed method

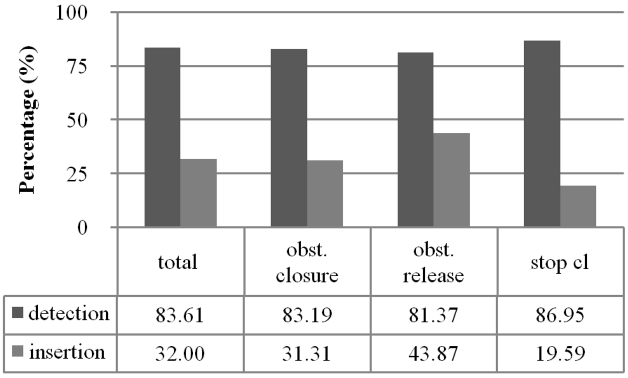
3.2 Decision rules

To reduce the spurious information in the candidates of landmarks, the decision rules are applied as in Fig. 3. There are three tests in the decision step: voicing test, silence test, and SCOG test. Decision rules are sequentially applied to the candidates.

The first step is a rejection of the peaks located in the voiced region. Ideally, most of obstruent consonant are located outside of the pairs of glottis landmarks marked with '+g' and '-g'. However, the voiced probabilities are measured for the candidates located outside of glottis landmarks in order to reduce the cumulative errors caused by glottis landmark detection. If the voiced probabilities are below 0.3 for the vicinity of candidates, the silence test is applied with candidates.

The second step is performed on the negative peaks in the candidates. It measures the RMS energy from the negative peaks. If the RMS energy has a value around 0-10dB during at least 60 ms, negative peaks of candidates are decided as stop closure landmarks. Stop closure is generated by the complete closure of oral cavity and there are significant energy drops. Therefore the stop closure has the lowest energy in the signal and it typically continues between 50 to 150 ms in duration [9]. For the remainder of candidates, SCOG is measured at the vicinity of peak locations to test the obstruent closure and release landmarks. The threshold of SCOG is determined based on the result of training database shown in Fig. 1. The landmarks of obstruent closure and release are selected if the SCOG measured around -candidates are above 1.5 kHz.

The final result with spurious candidates rejected are summarized as obstruent closure, obstruent release, and stop closure. Obstruent closure is a landmark associated with the onset time of fricative, affricate, and stop plosive. Obstruent release is related to the end of fricative and affricate. Thus, fricatives can be represented by the landmarks as obstruent closure and obstruent release. Affricates are represented as stop closure, obstruent closure and obstruent release. Stop consonant includes stop closure landmark and obstruent closure. Because stop plosive does not always placed with stop closure, the obstruent closure continuously occurred to stop closure is not determined as a stop burst.



	total	obst. closure	obst. release	stop cl
■ detection	83.61	83.19	81.37	86.95
■ insertion	32.00	31.31	43.87	19.59

Fig.4 Experimental results for obstruent consonant

4 Experimental results

Database is subdivided into development set and test set from the TIMIT corpus. The development set is composed of 100 sentences included in the sx groups of TIMIT training set. It is used to develop the landmark detection algorithm. For the performance evaluation, 840 sentences in the sx groups of TIMIT test set are selected. The numbers of tokens in the transcription of TIMIT test set are distributed as: fricative closure and release (3301), stop burst (3417), and stop closure (4068).

The performance evaluation is tested with 30 ms error boundary and hand correction for the test result is not included in the evaluation. Overall, 84% of landmarks are detected within 30ms and 32% of landmarks are inserted. The insertion rate includes both the neutral and insertion in the test. The results for each landmark are shown in Fig. 4. For the result of obstruent closure and release landmark, strident fricatives such as /s/, /sh/, /z/, and /zh/ are matched to the landmarks over 92%. 98% of landmarks are detected for the affricates. For the stop consonants, voiced stops have less performance than the unvoiced stops. Unvoiced stops are matched to the landmark over 90% and landmarks are detected with around 80% for voiced stops. Especially, the bilabial voiced stops such as /b/ have lowest detection rates with 71%. To summarize result, voiced fricatives and stops show lower performance than unvoiced consonants because the parameters of voiced consonants located in the vicinity of vowel are smoothed and the result of vowels affect the decision of voiced consonants. 86% of stop closure landmarks are detected by the test. Voiced stop closure has less performance than the unvoiced stop closure. The results of stop closure are influenced by the types of stops. Bilabial stop closure has the highest performance and alveolar stop closure has the lowest performance.

To compare with the earlier study, the result of burst landmark without hand-correction achieves to 81% of detection rates and 72% of insertion and neutral. The results of proposed algorithm have higher detection rates over 3% and lower insertion rates under 42% than the previous study.

5 Conclusion

The objective of this study focuses on the detection of obstruent consonant landmark. Landmark detection system is a first step in a knowledge-based speech recognition system. Therefore, accuracy of landmark detection is required to reduce the cumulative errors in the later process. The proposed algorithm suggests the refinement of landmark detection method. Previous method of landmark detection extracts the landmark by combining the extreme values in the sub-band energy. However the method based on the band energy profile can miss out the landmarks of obstruent consonant for CV or VC tokens. The changes of spectral center of gravity provide more effective information than the difference of sub-band energy. The overall result of experiment shows the improvement of landmark detection for the TIMIT corpus. Overall, 86% of landmarks are detected. The obstruent closure and release landmarks are detected with 83% and 82% detection rates, respectively. The stop closure landmarks are detected with 87% detection rate. The performance of proposed method has improved over the earlier landmark detection method.

Most of detection errors of proposed method are caused by the hard decisions at the final steps. Although the decision step uses the additional parameters, most of decision still determines the factors with hard decision. In order to improve the performance of landmark detection system, fixed threshold should be changed adaptively by updating the threshold from the training. Also, the various parameters used in decision step are required to make reliable decision.

References

[1] S. A. Liu, “Landmark detection for distinctive feature-based speech recognition,” *J. Acoust. Soc. Am.* Vol. 100, pp. 3417-3430, 1996.

[2] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. A. Liu, “Implementation of a model for lexical access based on features,” *in Proc. 6th Internat. Conf. Spoken Language Process.* Vol. 1, pp. 499-502, 1992.

[3] K. N. Stevens, “Toward a model for lexical access based on acoustic landmarks and distinctive features,” *J. Acoust. Soc. Am.* Vol. 111, pp. 1872-1891, 2002.

[4] R. Jakobson, G. Fant, M. Halle, *Preliminaries to speech analysis: the distinctive feature and their correlates*, MIT Press, Cambridge, MA, 1952.

[5] K. N. Stevens, *Acoustic Phonetics*, MIT Press, MA, 1998.

[6] A. M. Abdelatty Ali, “Acoustic-phonetic features for the automatic classification of fricatives,” *J. Acoust. Soc. Am.* Vol. 109, pp. 2217-2235, 2001.

[7] W. B. Kleijn and J. Haagen, *Speech Coding and Synthesis*, Elsevier Science, 1995.

[8] Wang, C. and S. Seneff, “Robust pitch tracking for prosodic modeling of telephone speech,” *in IEEE Proc. Internat. Conf. Acoustic, Speech, and Signal Process.* Vol. 3, pp. 1343-1346, Istanbul, Turkey, 2000.

[9] R. D. Kent and C. Read, *Acoustic Analysis of Speech 2nd edition*, Singular, 2002.

[10] P. Mermelstein, “Automatic segmentation of speech into syllabic units,” *J. Acoust. Soc. Am.* Vol. 58, pp. 880-883, 1975.