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

Acoustic-phonetic features for stop consonant place detection in clean and telephone speech

J.-W. Lee and J.-Y. Choi

Yonsei University, 134 Sinchon-dong, Seodaemun-gu, 120-749 Seoul, Republic of Korea
jaesuk2002@dsp.yonsei.ac.kr

This work classifies voiceless stop consonant place in CV tokens of English using burst release cues for clean (TIMIT) and telephone speech(NTIMIT). We compared the performance of cepstral coefficients to acoustic phonetics-motivated features such as center of gravity, burst amplitude and relative difference of formant amplitudes. In clean speech, cepstral coefficients resulted in better classification. However, for test data from NTIMIT, acoustic phonetic-based features outperformed cepstral coefficients, particularly if models were trained on clean speech. Augmenting cepstral coefficients with acoustic phonetic-based measurements resulted in the best performance only in clean speech. These findings suggest that cepstral coefficients are able to model speech in a given environment in finer detail, whereas acoustic phonetic-based features are more robust to changes in environment.

1 Introduction

Currently, the most popular method in speech recognition is a statistical method in which the speech recognizer models the pattern of speech signal sequences. Statistical methods have been successful since large training data are used to cover all the possible contextual variations. But problems arise when training data of some sounds are sparse. Also, because of heavy reliance on training data, the speech recognizer does not perform well if the operating environment does not match the training environment. Alternative speech recognition systems aim to overcome these problems by adopting acoustic phonetic-based features that reflect how the sounds are produced. These researches have also been classified as knowledge-based approaches.

The knowledge-based approach has been applied in the area of classification of stop consonants, which have distinct places of articulation [1]: at the labial (lips) for a /p/, alveolar (behind the teeth) for a /t/, and velar (at the velum) for a /k/. Spectral variation of these sounds is attributable to the fact that short noise burst is shaped by the resonance properties defined by a particular articulatory configuration [2].

A number of acoustic cues have been proposed which discriminate among /p/, /t/ and /k/ with spectral variation of a stop burst [3, 4, 5, 6, 7, 8]. And several experiments for better classification accuracy were demonstrated using combinations of acoustic cues which had been studied in the past [9, 10, 11, 12]. But, although those attempts yield high performance, the results of stop consonant place of articulation classification based on knowledge-based features does not outperform attempts based on spectral-based representation [13].

Studies which have been conducted on stop place detection have concentrated on finding new acoustic cues or obtaining higher accuracy. But research about considering changes in operating environment has not been investigated as much. In general, it is known that spectral-based representations are not robust to channel effects.

The objective of this paper is to classify voiceless stop consonant place in CV tokens of English using burst release cues for clean and spectrally impoverished speech. In particular, we will evaluate the robustness of acoustic-phonetic features to channel effects, compared to spectral-based representations.

2 Experiments

2.1 Database

The CV tokens are extracted using continuous speech from the TIMIT and NTIMIT database [14]. TIMIT database consists of 6300 utterances spoken by 630 speakers in quiet (4620 and 1680 utterances for training and test, respectively) and is labelled as words and phoneme unit. NTIMIT was collected by transmitting all 6300 original TIMIT utterances through various channels in the NYNEX telephone network and redigitizing them. NTIMIT database can be considered spectrally impoverished version of TIMIT database.

We used labelling of TIMIT and NTIMIT to extract the tokens which consist of closure, release burst and vowel. Flat sounds were excluded. Consequently, 1836 /p/, 3143 /t/ and 2905 /k/ tokens were collected as experimental samples from TIMIT and NTIMIT respectively.

2.2 Preparations

Release burst points and voice onset points of the following vowel are needed to conduct the classification of stop place. In this paper, we assume there is no error concerning release burst and voice onset point. For this purpose, points from TIMIT labelling are used. Energy of specific frequency range of the signal is calculated for the refining procedure of release burst and voice onset points, i.e., in 1.7 to 8 kHz and 60 to 400 Hz, respectively [15]. Then, the refined time points were selected where the rate of change of each band was maximal. But the range which can be modified from TIMIT labelling is restricted to 4ms.

A 512-point FFT was performed with a 6-ms Hanning window every 1ms. This is to obtain wideband spectrum and catch abrupt change of spectrum. An average power spectra technique also was used to obtain the spectrum of transient and burst. 15 spectra were averaged from release burst point onward if voice onset time (time from release burst to voice onset point) was longer than 20ms. If voice onset time was longer than 10ms, but shorter than 20ms, 8 spectra were averaged from release burst point onward. 3 spectra were averaged from release burst point onward if voice onset time was shorter than 10ms. This procedure is similar to Stevens et al., [11]; but this technique was empirically determined to be the best for excluding spectrum considered as silence.

2.3 Acoustic cues

There are a number of acoustic cues which have been proposed to classify stop place articulation. But there are limitations in selecting the acoustic cues to be used in this work, because spectrum of utterance in NTIMIT database is often absent about over 3.5 KHz. Based on this fact, it was decided to measure the following four acoustic cues: relative center of gravity, burst amplitude, voice onset time and Av-Amax23.

Acoustic analysis have shown that release burst of stop consonant in each group has different spectral peak: at low frequencies below 1KHz for /p/, at high frequency above 3KHz for /t/, and in the mid-frequency (1 to 4 KHz) for /k/ [4, 16]. Suchato has used this characteristic (concentration of energy in frequency domain) as an acoustic-phonetic feature [12]. The value of feature is center of gravity in the frequency unit of the power spectrum obtained from the time marked as the release burst to the time marked as the voicing onset of the following vowel. But in this work, center of gravity of each token is divided by center of gravity of sentence that each token is extracted. For this reason, we call it relative center of gravity.

Burst amplitude also reflects stop place articulation. Alveolar stop is the strongest burst, and labial stop is the weakest burst. Zue [3] and Edward [8] measured burst amplitude as the ratio of the maximum root mean square (RMS) amplitude of the following vowel to the RMS amplitude of the burst. This measurement was adopted in this work. RMS amplitude of the burst was calculated using an averaged power spectrum which.

Voice onset time (VOT) can also be a cue to place of articulation of the consonant [3, 6, 8, 17]. VOT is the time between release burst points and voice onset points of the following vowel. The general knowledge is that labial has the shortest VOT, while velar stop has the longest VOT. This temporal feature can be an important cue when information of spectrum is impoverished.

Finally, Av-Amax23 was measured. Av-Amax23 is the log of the ratio of the amplitude of the first formant prominence measured at voicing onset to the maximum amplitude of the spectrum in the F2 and F3 range at release burst [11, 12]. Value of feature is expected to be least for velars, and highest for labials, because velars have spectral peak in mid-frequency (1 to 4 KHz) and labial stop is the weakest burst. Maximum amplitude of the spectrum in the F2 and F3 range at release burst used an averaged power spectrum which was calculated in advance.

The values of acoustic parameters, relative center of gravity, burst amplitude, voice onset time and Av-Amax23, were measured to form one feature vector as the representation of a stop consonant place. Then each place was modeled by mixtures of 4 Gaussian densities. The performance according to the probability of correct classification was achieved by the Bayes classifier.

2.4 Cepstral Coefficients

The even component of the complex cepstrum, i.e., real cepstrum, was used to compare the robustness of acoustic phonetic features against channel effect. The cepstral coefficients were computed over the original frequency range (0~8 KHz) of burst release. Cepstral mean subtraction (CMS) was also conducted. It has been well known that mean subtraction improves the performance of a system in which training is done on one channel condition while testing is done on another channel condition [18].

3 Results

3.1 Statistical analysis

The measurements obtained for all voiceless stop consonant tokens in the training subset of TIMIT were examined using an analysis of variance (ANOVA). In this test, significance level (α) was 0.01. If p-value corresponding to the F-ratio at the right degrees of freedom is smaller than α, we may consider mean difference of three voiceless stop consonant place (labial, alveolar and velar) is significant and is originated from the place effect rather than the error. Even though p-value is smaller than α, it does not mean that all of the mean differences between groups are significant. The pair-wise F-ratio was used on the three possible pairs to test the significance of the mean difference between two groups. The F-ratio is listed in Table 1. The degree freedom between groups is 2 for three groups, 1 for pair-wise. Within groups is 5906 for three groups, 3697 for labial vs. alveolar, 3526 for labial vs. velar, and 4589 for alveolar vs. velar.

		Rela. C.O.G	Burst amp.	V.O.T	Av-Amax23
F-ratio	Three groups	2065.8 (0)	1117.4 (0)	324.8 (0)	662.9 (0)
	Labial vs. Alveolar	4040.0 (0)	1994.7 (0)	127.2 (0)	616.7 (0)
	Labial vs. Velar	1147.6 (0)	621.2 (0)	621.0 (0)	1315.0 (0)
	Alveolar vs. Velar	1214.3 (0)	176.1 (0)	83.2 (0)	101.9 (0)

Table 1 ANOVA results (F-ratio and p-value) among Labial (/p/), alveolar (/t/) and velar (/k/) for 4 acoustic phonetic features in the training subset of TIMIT. P-values are shown in parentheses.

Table 1 shows that p-values of all groups are 0s, which means each acoustic phonetic feature is originated from different distributions. Relative center of gravity shows the biggest F-ratio in the measurement of three groups. It also has quite big F-ratio values in the pair-wise. From this fact, we can consider relative center of gravity is the best feature among four features. F-ratio between alveolar and velar is mostly smaller than other pair-wise groups. Voice onset time shows worst significant difference. But we can expect its contribution to the classification is still significant.

3.2 Place detection

The detection rate of each acoustic phonetic feature for stop consonant place detection is shown in Fig. 1 when the training and test circumstances are changed or not.

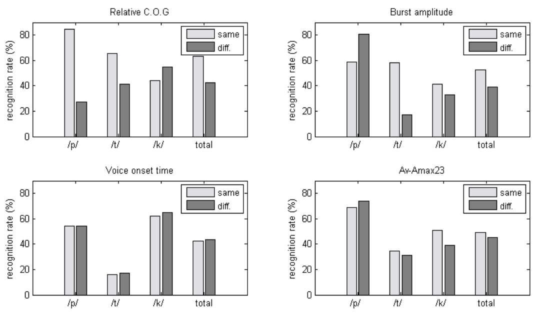

Fig. 1 Detection rate of relative center of gravity, burst amplitude, voice onset time and Av-Amax23 for stop consonant place detection. Bright and dark bars are when the environment of training and test is same or different, respectively.

As we could expect in Table 1, recognition rate in same circumstance gets better when F-ratio is bigger. Fig. 1 shows that performance of most features becomes worse when circumstance of training and test are different, especially classification rate for relative center of gravity because of huge spectrum distortion. But performance of voice onset time was rarely influenced by the environment change. This shows that feature related to duration is robust to change of circumstance.Av-Amax23 was also quite good feature which shows little decrease in detection rate.

Acoustic phonetic features and cepstral coefficients were next used to detect stop consonant place as training and test environment is changed or not. Fig. 2 shows the best classification cases for combination of acoustic phonetic features and cepstral coefficients, respectively.

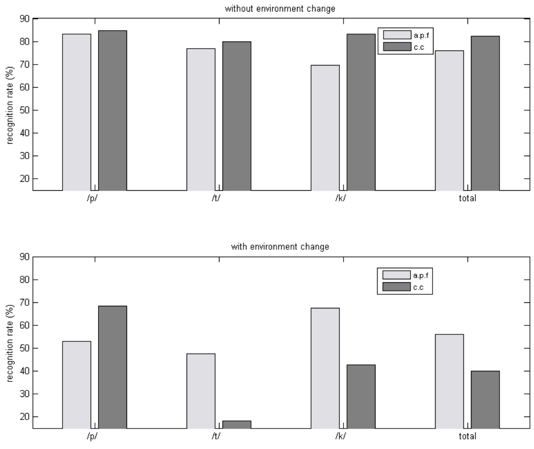

Fig. 2 The best detection cases of cepstral coefficients and combination of acoustic phonetic features. All acoustic cues are used when environment for training and test is not changed. 3 acoustic cues except burst amplitude are used for changed environment. 10 and 4 order cepstral coefficients are used for not changed and changed environment, respectively.

When training and test circumstance is not changed, performance of cepstral coefficient was better than one of acoustic phonetic features in all places. Total classification rate was achieved a probability of correct classification equal to 82.3% for cepstral coefficients and 76% for acoustic phonetic features. Bottom figure of Fig. 2 shows classification rate decreases overall when model is trained in clean speech and tested in telephone speech. In this case, detection rate of labial place for cepstral coefficients only is higher than one of acoustic phonetic features. Especially, detection rate of alveolar place (18.2%) using cepstral coefficients shows no ability to classify in that place. Totally, acoustic phonetic features (55.95%) outperform cepstral coefficients (39.95%). This is an indication that acoustic phonetic features are more robust than cepstral coefficients when environment is changed.

Additional experiment, combining both types of measurements, was conducted. Combination feature leads to the best performance, 89.3%, 84%, 84.8% and 85.7% for labial, alveolar, velar and total classification rate, respectively, when model is trained and tested in clean speech. It means that acoustic phonetic measurements provide complementary information to conventional cepstral coefficient measurements. But there was no enhancement under different environment of training and test. It may be that poor performance of cepstral coefficients does not supply any additional information to classification.

4 Conclusion

In this paper, cepstral coefficients and acoustic phonetic features such as relative center of gravity, burst amplitude, voice onset time and Av-Amax23 were examined in detection of voiceless stop consonant place in CV tokens of English for clean (TIMIT) and telephone speech (NTIMIT). When the environment for training and test is not changed, cepstral coefficients resulted in better classification. But performance of acoustic phonetic features were much better than cepstral coefficients under different environment for training and test, which means acoustic phonetic features are more robust to changes in environment.

In this paper, only four acoustic features are used due to band-limited spectrum of telephone speech. Further studies will focus on extracting more features from telephone speech to enhance the detection performance.

References

[1] P. Lieberman and S. Blumstein, *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge, U. K.: Cambridge University Press (1988)

[2] R. D. Kent and C. Read. *The Acoustic Analysis of Speech*, Singular, San Diego (1992)

[3] V. W. Zue, “Acoustic characteristics of stop consonants: A controlled study,” D.Sc. dissertation, Mass. Inst. Technol., Cambridge (1979)

[4] K. N. Stevens and S. Blumstein, “Invariant cues for place of articulation in stop consonants”, *J. Acoust. Soc. Am.* 64, 1358-1368 (1978)

[5] S. Blumstein and K. N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants", *J. Acoust. Soc. Am.* Vol. 66, no. 4, pp. 1001-1017 (1979)

[6] Kewley-Port, D. "Time-varying features as correlates of place of articulation in stop consonants", *J. Acoust. Soc. Am.* 73, 322-355 (1983)

[7] Kewley-Port, D., and Luce, P. A. "Time-varying features of initial stop consonants in auditory running spectra: A first report", *Perception & Psychophysics*, 35, 353-360 (1984)

[8] T. J. Edwards, "Multiple features analysis of intervocalic English plosives", *J. Acoust. Soc. Am.* Vol. 69, no. 2, pp. 535-547 (1981)

[9] Repp, B. H., and Lin, H. "Acoustic properties and perception of stop consonant release transients", *J. Acoust. Soc. Am.* 85, 379-396 (1989)

[10] M. Hasegawa-Johnson. "Formant and burst spectral measurements with quantitative error models for speech sound classification", Ph.D. thesis, Massachusetts Institute of Technology. (1996)

[11] K. N. Stevens and S. Y. Manuel, "Revisiting place of articulation measures for stop consonants: Implications for models of consonant production", *International Congress on Phonetic Sciences, 2*, 1117-1120 (1999)

[12] Suchato, A. "Classification of stop consonant place of articulation", Ph.D. thesis, Massachusetts Institute of Technology. (2004)

[13] Halberstadt, A. "Heterogeneous acoustic measurements and multiple classifiers for speech recognition", Ph.D. thesis, Massachusetts Institute of Technology. (1998)

[14] S. Seneff and V. Zue, "Transcription and alignment of the TIMIT database", in *Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*, J. S. Garofolo, Ed. Gaithersburg, MD: NIST (1988)

[15] Liu, S. A. "Landmark detection for distinctive feature-based speech recognition", *J. Acoust. Soc. Am.* 100, 3417-3430 (1996)

[16] Halle, M., Hughes, G. W., and Radley, J.P. "Acoustic properties of stop consonants", *J. Acoust. Soc. Am.* 29, 107-116 (1957)

[17] Umeda, N. "Consonant duration in American English", *J. Acoust. Soc. Am.* 61, 846-858 (1977)

[18] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker Recognition: A Feature-based Approach", IEEE Signal Processing Mag., vol. 13, no. 5, pp. 58-71 (1996)