AUTOMATIC DETECTION AND CLASSIFICATION OF STOP CONSONANTS USING AN ACOUSTIC-PHONETIC FEATURE-BASED SYSTEM

Ahmed M. Abdelatty Ali*, Jan Van der Spiegel* and Paul Mueller♣

**Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104-6390, USA, ♣Corticon, Inc., 155 Hughes Rd, King of Prussia, PA 1940, USA.*

ABSTRACT

A new acoustic-phonetic feature-and knowledge-based approach for the detection and classification of stop consonants in speaker-independent continuous speech is proposed. A system is built which automatically extracts stop consonants from continuous speech. The detected stop consonants are then passed to the classification system, which classifies them according to their voicing and place of articulation. The system combines multiple static and dynamic features in a multi-level decision process. It uses an auditory-based front-end processing and incorporates new algorithms for the extraction and manipulation of the acoustic-phonetic features that proved to be useful in the recognition process. The system is tested on 300 sentences from 30 different speakers of the TIMIT database with 6 different dialects. The stops were detected with an accuracy of 88% (8% substitution, 3% insertion and 1% deletion errors). The classification accuracy was 86%, with 97% voicing accuracy and 90% place of articulation accuracy.

1. INTRODUCTION

Despite the long history of research on the acoustic characteristics of stop consonants, current state-of-the-art automatic speech recognition (ASR) systems are still incapable of performing accurate fine phoneme distinctions for this class of sounds (and for some other classes as well). One of the main reasons for this is the dynamic, short, speaker- and context-dependent nature of these sounds. The information that exists in the literature is not sufficient or consistent enough to be integrated in an ASR system.

In this work we concentrate on the acoustic-phonetic characteristics responsible for detecting and classifying the stop consonants. The detection (extraction) of the stop consonants is performed as part of a segmentation and coarse categorization system [1]. The segmentation of speech is performed first by detecting the boundaries between, and identifying, three classes: silences, sonorants and obstruents. The obstruents are further categorized into: fricatives, stops and affricates. The output of the system is 4 main classes: fricatives (and affricates), stops, sonorants, and silences. The three main features that distinguish this approach are, first, the combination of explicit and implicit segmentation; second, allowing overlapping categories; and third, its ability to accurately discriminate between stops and fricatives in a knowledge-based manner. The classification of the detected stop consonants is then performed by detecting the place of articulation and voicing [2, 3].

2. FRONT-END PROCESSING

The front-end signal processing that is used in our system is a biologically oriented filter-bank system. It is based on the system developed by Seneff and described in detail in [13]. The block diagram is given in Fig. (1). The system gives two outputs, namely the mean-rate output and the Generalized Synchrony Detector (GSD) output. The front-end Bark-scaled filter bank consists of 36 filters with 20dB/decade high frequency pre-emphasis. The reasons for choosing this system are described in detail in [1].

The Hair Cell Synapse Model block includes the short-term adaptation and forward masking effects, which play a useful role in the segmentation process. By exponentially attenuating the response to steady-state signals and enhancing abrupt changes, it sharply marks the boundaries between different phoneme types. This in turn leads to larger boundary signal peaks and hence better segmentation. This block is also known to improve the immunity of the system to noise [13].

3. THE ALGORITHM

The overall system is shown in Fig. (2). The segmentation and categorization algorithm starts by detecting silences including closures of stop consonants. The criterion used for detecting silences is based on the presence of very low total energy (below a very tight threshold) in the outputs of both the synchrony and mean-rate *or* the presence of low total energy (below a more relaxed threshold) *and* low energy in the low- and high-frequency ranges. The low frequency range is defined here as the range below 400Hz, while the high frequency range is defined as the range above 3.5KHz. All energy quantities are normalized with respect to the maximum energy of the utterance.

After silence detection, sonorants are extracted. The periodic nature of the sonorants, and the sensitivity of the generalized synchrony detector to periodicity, make this step straightforward and simple. Sonorants are detected by the presence of any of the following:

- the sum of the synchrony outputs below 500Hz is greater than a certain empirically determined threshold (*sonorant_threshold_1*), *or*

- the sum of the synchrony outputs below 500Hz is greater than another, smaller, empirically determined threshold (*sonorant_threshold_2*) and the total synchrony energy is relatively small (below a certain threshold) and the largest spectral peak lies below 1KHz, *or*
- the largest spectral peak lies below 500Hz, *or*
- average zero crossing rate is below a certain threshold.

The above 4 conditions are applied only to sounds that were not detected as silences. They are empirically designed to cover all possible variations of sonorants, i.e. strong vowels, weak unstressed vowels, semivowels and nasals.

Till this step, no explicit segmentation was performed. All the processing was performed to locate *regions* of silences and sonorants. No boundaries have been explicitly detected yet. The next step extracts the boundaries in the speech signal. It locates those boundaries using smoothed derivatives of some key features. Since we are interested only in boundaries which are in, or on the edges of, the non-sonorant regions, suitable thresholds could be obtained for performing accurate segmentation for those regions. Thus the segmentation step performs a smoothed derivative for the following features:

- the energy in the high frequency region (above 3.5KHz),
- the energy in the low frequency region (below 800Hz),
- the energy in the middle frequency region (between 500Hz and 2KHz),
- the spectral center of gravity and the maximum energy value.

All those features are obtained from the mean-rate outputs. To avoid the effect of small variations, only variations above certain thresholds are passed by applying *center clipping* to values below the thresholds. Those thresholds are designed as functions of the mean values of the feature variation function to avoid any dependence on their absolute values. After proper normalization, the contributions from the different features are added up to form the boundary signal which describes the rate of rise (RR), or rate of change, of the combined, previously mentioned, acoustic features. This signal, (call it the RR signal), consists of peaks that mark the positions of significant changes in the spectrogram and hence possible segment boundaries.

Obstruents are then extracted, from the regions, which are neither silences nor sonorants, by satisfying a static *or* a dynamic condition. The static condition is that the spectral center of gravity or zero crossing rate being above certain thresholds, which indicates the presence of high frequency noise. The dynamic condition is that the RR signal is above a certain threshold in the area around the region, subject to some continuity constraints. This condition captures sudden changes characteristic of stops and affricates.

Thus, the speech signal is classified into silences, sonorants and obstruents. Regions which did not fall under any of those regions will be categorized using some continuity and timing constraints with the neighboring segments. Moreover, we also have a boundary signal (RR) which consists of peaks that mark the boundaries between the different segments. It is worth mentioning that all the thresholds used in the algorithm are relative to either the mean or maximum values. And hence they are insensitive to any variation in the intensity of speech. They are also independent of inter- and intra-speaker variations. This makes the algorithm suitable for speaker-independent automatic recognition.

The next step is to divide the obstruents into fricative and stops. Fricatives are characterized by longer duration and/or a more gradual rise time compared to the shorter and sharply rising stop consonants. Moreover, the release of the stops is usually preceded by a silence period due to the stop closure. An exception to this property is the flap which has unique properties and could be easily extracted separately [3]. This condition however (i.e. the presence of a closure period) is a necessary but not sufficient condition for the rest of the stops. Many fricatives have a preceding period of silence either because of their presence in the beginning of a word/utterance, or the short silence period that sometimes separates the fricative from the neighboring vowel.

Another property which characterizes the stops is the sudden drop in the total energy shortly following the sharp rise. This could be because of the period of aspiration which usually has less energy than the initial release, or the short period of the stop which leads to a sudden drop in the energy. This sudden drop does not exist in fricatives which are characterized by a slowly rising or steady energy level.

The main parameter used for distinguishing between stops and fricatives is the Rate-of-Rise-to-Duration Ratio (RRDR). This parameter integrates the rate of rise and the duration in one value. Since stops are characterized by shorter duration and faster rise time, it is expected that the RRDR would be large for stops compared to fricatives. The rate of rise is found using the boundary signal (RR) described before where the rate of rise is represented by its peak value in the neighborhood of the release. Moreover, the non-continuant nature of the stops puts a limit on the maximum duration it could have. An upper limit was found to be 125ms, any obstruent having a duration larger than 125ms is assumed to be a non-stop. On the other hand, a lower limit for fricatives was found to be 18ms below which the sound is assumed non-fricative.

To summarize, a stop is detected if the sound is an obstruent with duration less than 125ms *and* any of the following is satisfied:

- the release duration is less than 18ms, *or*
- a preceding closure exists and RRDR is larger than a certain threshold, *or*
- a preceding closure exists and a sudden drop in energy exists shortly after the release.

On the other hand, a fricative is detected if the sound is an obstruent with duration larger than 18ms *and* any of the following is satisfied:

- the duration is larger than 125ms, *or*
- there is no preceding closure, *or*
- the RRDR is lower than a certain threshold, *or*
- the sound is not a stop as described above.

The RRDR is also useful in extracting affricates which include /tcl ch/ and /dcl jh/ as well as other combinations which do not formally exist linguistically but occur often in continuous speech like: /ks/ in “dark suit”. Those affricates are characterized by a sharp rise and a long duration. This causes their RRDR value to be in the medium range. Therefore sounds which have an RRDR below the threshold for stops but above the threshold for fricatives are assumed to be affricates.

The detected stop consonants are then passed to the classification system which classifies the stops according to their voicing and place of articulation [2,3]. The voicing detection algorithm used the following three features:

1. Voicing during closure (prevoicing)

2. Voicing onset time

3. Closure duration.

The place of articulation detection algorithm used the following features:

1. Burst frequency

2. The second formant of the following vowel

3. The maximum normalized spectral slope

4. The burst frequency prominence

5. Formant transitions before and after the stop

6. The voicing decision.

The details of the algorithms, which use the above features to perform voicing and place of articulation detection, are described in [2,3].

4. RESULTS

The segmentation algorithm was designed using continuous speech from 6 speakers in the TIMIT database and tested on continuous speech of 30 different speakers from 6 different dialects from the TIMIT database which contained more than 5000 phonemes. The accuracy was 92%, with 4% substitution, 3% insertion and 1% deletion. The stops were detected with an accuracy of 88% (8% substitution, 3% insertion and 1% deletion errors). The majority of the insertion errors are stop consonants erroneously inserted after, or before, a silence. The confusion matrix for the substitution errors is shown in table (1).

	Detected as Stops	Detected as Fricatives	Detected as Sonorants	Detected as Silences
Stops	92%	6%	0%	2%
Fricatives	10%	86%	4%	0%
Sonorants	0.5%	0.5%	97%	2%
Silences	2%	2%	2%	94%

Table (1). Confusion matrix for the segmentation and categorization system. This table shows the substitution errors only. Substitution accuracy is 96%.

The classification results are shown in tables (2), (3) and (4). Table (2) shows the voicing detection results which have an accuracy of 97%. Table (3) shows the place of articulation detection accuracy of 90%, while table (4) shows the overall classification accuracy of 86%.

	Detected as voiced	Detected as unvoiced
Voiced	95%	5%
Unvoiced	2%	98%

Table (2). Confusion matrix for voicing detection on 1200 stops from 30 speakers. Recognition accuracy is 97%.

	Detected as alveolar /t,d/	Detected as velar /k,g/	Detected as labial /p,b/	Detected as flap /dx/
Alveolar	91%	6%	3%	0%
Velar	3%	88%	9%	0%
Labial	6%	6%	86%	2%
Flap	X	2%	4%	94%

Table (3). Confusion matrix for the place of articulation detection. ‘X’ denotes a “don’t care”, since a flap could be either a /t/ or a /d/. Recognition accuracy is 90%.

	Det. as t	Det. as d	Det. as k	Det. as g	Det. as p	Det. as b	Det. as dx
t	88%	3.5%	5%	0%	3.5%	0%	0%
d	3%	89%	0%	7%	0%	1%	0%
k	2.5%	0%	86.5%	1%	10%	0%	0%
g	2%	2%	10%	76%	0	10%	0%
p	7%	0%	7%	0%	85%	1%	0%
b	0%	5%	0%	5%	2.5%	86%	1.5%
dx	X	X	0%	2%	0%	4%	94%

Table (4). Confusion matrix for the overall recognition of 1200 stops, for 30 speakers from 6 different dialects, using the voicing and place of articulation detection techniques. Overall accuracy is 86%.

5. CONCLUSION

An acoustic-phonetic, knowledge-based, system for the extraction and classification of stop consonants in continuous speech is built. It categorizes the speech into fricatives, stops, sonorants (vowels and semi-vowels) and silences. It then classifies the detected stops based on their voicing and place of articulation. The system uses multiple features (static and dynamic), allows overlapping phoneme categories, minimizes information loss, and integrates explicit and implicit segmentation techniques. It achieves an 88% detection accuracy and an 86% classification accuracy when tested on continuous speech from 30 speakers of the TIMIT database not used in the design process.

ACKNOWLEDGMENTS

This work is supported by a grant from Catalyst Foundation.

REFERENCES

[1] Ali, A.M.A., "Feature-based segmentation and categorization of continuous speech", Technical Report, TR-CST25JUL98, Center for Sensor Technologies, University of Pennsylvania, 1998.

[2] Ali, A.M.A., *et al.*, "Acoustic-phonetic features for the automatic recognition of stop consonants", *J. Acoust. Soc. Am.*, pp. 2777-2778, 103 (5), 1998.

[3] Ali, A.M.A., "Acoustic-phonetic features for the automatic recognition of stop consonants", Technical Report, TR-CST22DEC1997, Center for Sensor Technologies, University of Pennsylvania, 1997.

[4] Ali, A.M.A., *et al.*, "An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants", *Proc. IEEE ICASSP-98*, vol.(II), pp. 961-964, 1998.

[5] Glass, J.R., "Finding acoustic regularities in speech: applications to phonetic recognition", Ph.D. thesis, MIT, 1988.

[6] Glass, J.R., and Zue, V.W., "Multi-level acoustic segmentation of continuous speech", *Proc. ICASSP-88*, pp. 429-432, 1988.

[7] Hatazaki, K., et al., "Phoneme segmentation using spectrogram reading knowledge", *Proc. ICASSP-89*, pp. 393-396, 1989.

[8] Leung, H.C., and Zue, V.W., "A procedure for automatic alignment of phonetic transcriptions with continuous speech", *Proc. ICASSP-84*, pp. 2.7.1-2.7.4, 1984.

[9] Liu, S.A., "Landmark detection for distinctive feature-based speech recognition", *J. Acoust. Soc. Am.*, pp. 3417-3430, 100 (5), 1996.

[10] Mermelstein, P., "Automatic segmentation of speech into syllabic units", *J. Acoust. Soc. Am.*, pp. 880-883, 58 (4), 1975.

[11] Reddy, D.R., "Phoneme grouping for speech recognition", *J. Acoust. Soc. Am.*, pp. 1295-1300, 41 (5), 1967.

[12] Seneff, S., "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing", *J. of Phonetics*, 16, pp. 55-76, 1988.

[13] Seneff, S., "Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model", PhD Dissertation, MIT, 1985.

[14] Stevens, K.N., "Acoustic correlates of some phonetic categories", *J. Acoust. Soc. Am.*, pp. 836-842, 68 (3), 1980.

[15] Zue, V.W., et al., "Acoustic segmentation and phonetic classification in the SUMMIT system", *Proc. ICASSP-89*, pp. 389-392, 1989.

[16] Zwicker, E., "Subdivision of the Audible Frequency Range into Critical Bands (Fequenzgrupper)", *J. Acoust. Soc. Am.*, 33, 1961, 248.

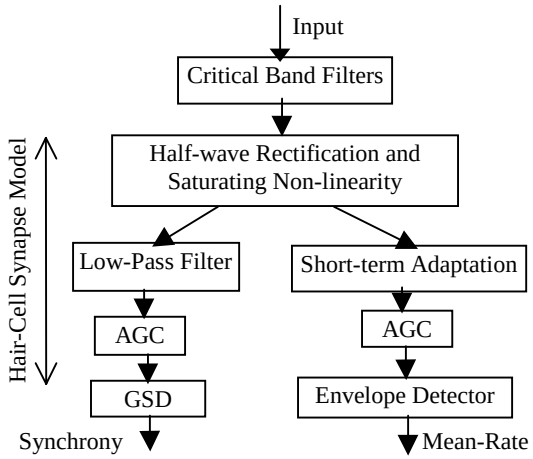

Fig. (1) Block diagram of the auditory-based front-end processing system.

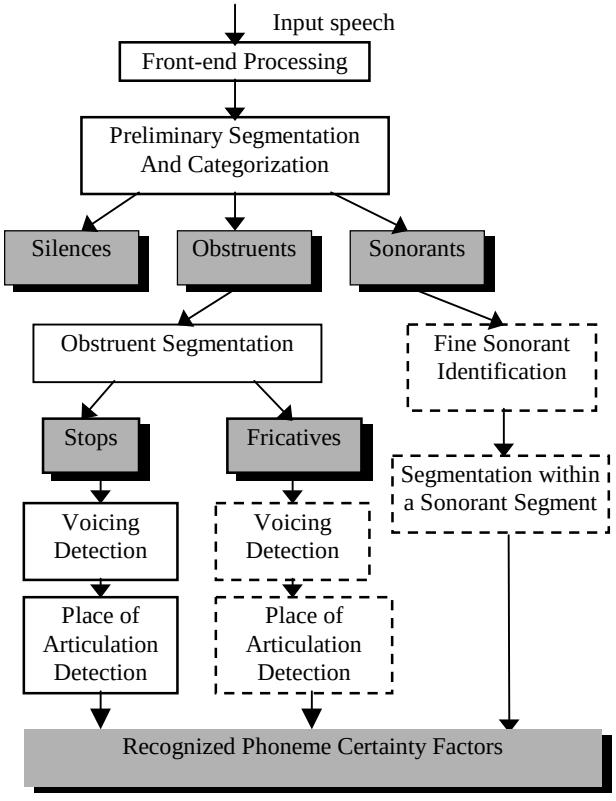

Fig. (2) Block diagram of the overall system.