Acoustic–Phonetic Features for the Automatic Classification of Stop Consonants

Ahmed M. Abdelatty Ali, *Member, IEEE*, Jan Van der Spiegel, *Senior Member, IEEE*, and Paul Mueller

***Abstract*—In this paper, the acoustic–phonetic characteristics of the American English stop consonants are investigated. Features studied in the literature are evaluated for their information content and new features are proposed. A statistically guided, knowledge-based, acoustic–phonetic system for the automatic classification of stops, in speaker independent continuous speech, is proposed. The system uses a new auditory-based front-end processing and incorporates new algorithms for the extraction and manipulation of the acoustic–phonetic features that proved to be rich in their information content. Recognition experiments are performed using hard decision algorithms on stops extracted from the TIMIT database continuous speech of 60 speakers (not used in the design process) from seven different dialects of American English. An accuracy of 96% is obtained for voicing detection, 90% for place of articulation detection and 86% for the overall classification of stops.**

***Index Terms*—Acoustic–phonetic, feature extraction, phoneme recognition, speech recognition, stop consonants.**

NOMENCLATURE

ALSD	Average localized synchrony detector developed by the authors [3], [5].
Burst Spectrum	Spectral shape during the burst (i.e., release) of the stop.
BF	Burst frequency defined in (1).
DRHF	Dominance relative to the highest filters defined in (4).
Fi	ith formant.
GSD	Generalized synchrony detector developed by Seneff [36], [37].
LINP	Laterally Inhibited MDP defined in (5).
MNSS	Maximum normalized spectral slope defined in (2).
MDP	Most dominant peak defined as the peak with the largest amplitude or slope.
Prevoicing	Voicing during the closure period of the stop.
VOT	Voicing onset time: time from the stop release to the voicing onset of the following vowel.
VF2	Second formant of the following vowel.

Manuscript received June 29, 1999; revised August 15, 2001. This work was supported under a grant from the Catalyst Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rafid A. Sukkar.

A. M. A. Ali is with Texas Instruments, Inc., Research and Development, Warren, NJ 07059 USA and also with the Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104-6390 USA (e-mail: ahm@ee.upenn.edu).

J. Van der Spiegel is with the Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104-6390 USA (e-mail: jan@ee.upenn.edu).

P. Mueller is with Corticon, Inc., King of Prussia, PA 19406 USA (e-mail: cortion@aol.com).

Publisher Item Identifier S 1063-6676(01)09663-8.

I. 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. 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 neither sufficient nor consistent enough to be integrated in an ASR system.

The stop consonants, /t/ /k/ and /p/ and their voiced cognates /d/ /g/ and /b/, are a class of sounds which is formed by the greatest degree of obstruction and a complex of movements in the vocal tract. The articulators form an oral occlusion (closure) behind which pressure is built up. The location of the oral occlusion, i.e., the place of articulation, could be bilabial (/p/ and /b/), alveolar (/t/ and /d/) or palatal/velar (/k/ and /g/). During the closure period, the vocal cords may or may not vibrate. If they do, the stop is said to be prevoiced. After the closure phase comes the release phase. In the release, the oral occlusion is broken, releasing the air pressure and allowing the air to resume its flow. When stops are released, an audible burst of noise results. This burst of noise is different from the fricative noise in being transient and not prolongable. This gives the stops the property of not being continuants.

In this work, we investigate the acoustic-phonetic characteristics of stop consonants. We combine expert knowledge and statistical analysis in a hybrid approach, to gain a better understanding of the role of various static and dynamic features in the recognition process (individually and combined). The designed system can be best described as a statistically guided knowledge-based system. It uses a new auditory-based front-end that generates mean-rate and synchrony outputs.

We concentrate here on the characteristics responsible for classifying the stops (i.e., detecting the place of articulation and voicing). Extracting the stops, on the other hand, is discussed in more detail elsewhere [5] as a part of a segmentation and phoneme categorization system. This system is used in the present experiments to extract the stops and mark their different segment boundaries.

In the next section, the acoustic–phonetic features of stop consonants, which exist in the literature, are discussed. In the

following sections, the results of our research on the characteristics of stop consonants and their automatic classification are discussed.

II. Acoustic–Phonetic Features

A. Formant Transitions and Burst Spectrum

The role of the formant transitions and the burst spectrum, and their relative importance, has been considerably researched and debated in the literature. However, despite the wealth of information that exists in the literature, a considerable amount of ambiguity and contradiction also exists. A comprehensive survey of previous research [5] illustrates that the burst spectrum and the formant transitions are important for the place of articulation detection. Their role, however, in the voicing detection seems to be secondary. These two features are actually closely related, functionally equivalent and complementary [20]. Their *perceptual weight* seems to depend on the degree of their salience, such that the formant transition role becomes more significant when the transitions involved are sharp and clear. On the other hand, its role (perceptual weight) becomes negligible when the transition is slight and ambiguous.

It is also clear after this long history of research that absolute acoustic invariance is not possible for the stop place detection. *Relational invariance*, where the feature depends on the neighboring vowel in a well-defined manner, on the other hand, seems to be a more plausible and useful approach.

B. Burst Amplitude

Previous research found that labial stops are usually weaker than alveolars and velars [21], [23], [45]. Perceptual experiments on syllable-initial alveolar and labial stop consonants also showed that the relative amplitude of the burst can influence the identification of alveolar and labial place of articulation [30]. This influence is more profound for voiceless than for voiced stops and is evident only for stops which have ambiguous spectra. This indicated that the amplitude does play a role in the place detection. This role however seems to be secondary, since the burst spectrum seems to override its effect.

C. Durations and Voicing

The stop consonant consists of a closure interval, a release (transient, frication and aspiration) interval and a transition interval (from voicing onset to the vowel's nucleus). The durations of these different segments, and of the stops as a whole, were investigated by many researchers [6], [14], [21], [27], [42], [43], [45]. It was found that the mean voicing onset time (VOT) for voiceless stops is longer than their voiced cognates. It was also clear that the VOT could play a major role in the voicing detection but not in the place of articulation detection, in which its role is secondary at best. Moreover, in spite of its importance in voicing detection, it is not expected to be able to perform the task alone. Other features are needed to resolve the significant overlap that exists between voiced and unvoiced VOT distributions especially for stops in different contexts.

Another feature that was investigated by many researchers is the presence of voicing during the stop closure (prevoicing). This feature is found to be a sufficient, but not necessary, condition for voicing.

III. Acoustic–Phonetic Classification

In this section, the experiments performed on the stop consonants for the place of articulation and voicing detection are discussed. We made use of the wealth of information that exists in the literature, our own acoustic and spectrogram reading knowledge and different statistical tools to build the resulting "knowledge-based" system. Statistical discriminant analysis, histogram analysis, information transmission analysis and decision trees are some of the statistical tools that helped design the system (i.e., to determine thresholds, evaluate features, combine features, etc.).

This system is designed using continuous speech from ten speakers (five males and five females) from the TIMIT database and then tested on 1200 stops extracted from continuous speech of 60 different speakers (not used in the design phase) from seven different dialects of American English in the TIMIT database. We will concentrate here on the place and voicing detection, the stop manner of articulation is detected by another system developed to segment and categorize the phonemes in an utterance [5]. Since the experiments discussed in this work report the classification results, errors obtained in the detection and segmentation were excluded from the results. The output of the segmentation system marks the closure and release segments of the stop. It also marks the point of voicing onset as evidenced by the presence of low frequency energy in the F0 and F1 regions. This is explained schematically in Fig. 1.

The front-end signal processing system that is used in our experiments is an auditory-based Bark-scaled filter-bank system. It is a modification to the system developed by Seneff and described in detail in [36], [37]. The block diagram is given in Fig. 2. The filter bank used is a bank of 36 critical-band filters (Bark scale) with the distribution given by Zwicker [46]. It is preceded by a 20 dB/decade high-frequency pre-emphasis. This, and other, auditory-based distributions have proved to yield better results in ASR applications [7], [15], [24], [25], [35], [38]. The system gives two outputs, the mean rate and the synchrony output. The synchrony describes the temporal pattern and is extracted using the average localized synchrony detector (ALSD) [3], [5]. This is a modification to Seneff's generalized synchrony detector (GSD) [36], [37], developed by the authors to alleviate some of its limitations. Mainly it employs a novel spacial averaging technique to enhance its formant extraction ability while suppressing the spurious peaks. The synchrony is used for its superior formant extraction ability, higher response to periodic signals and higher immunity to noise, while the mean rate is used for its higher sensitivity and better ability in describing the overall spectral shape. This is in agreement with auditory neurobiology, where the average response and the temporal pattern of the neural firings play complementary rules that are similar to the rules employed in this work [17], [18].

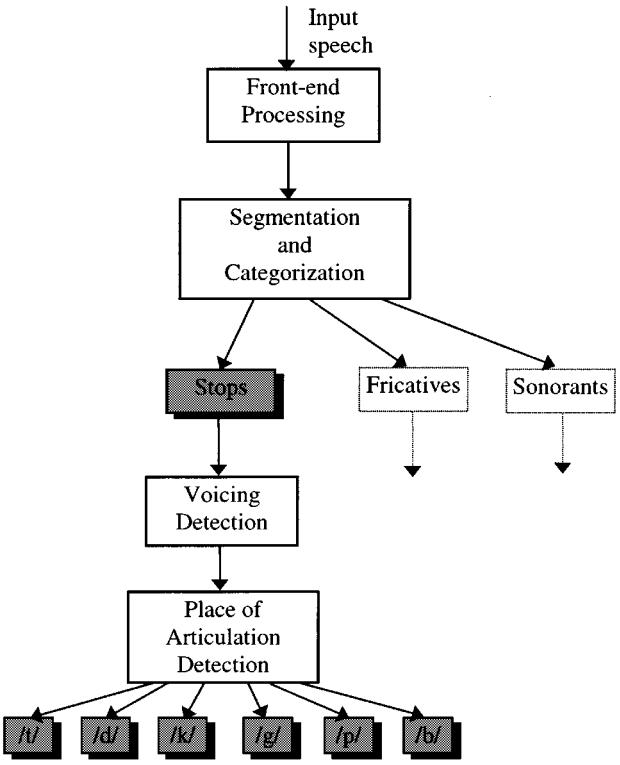

Fig. 1. Block diagram of the stop recognition system used.

A. Voicing Detection

Three main features were found to be useful for voicing detection:

1) voicing during closure (prevoicing);
2) voicing onset time (VOT);
3) closure duration.

These three features are combined in the algorithm shown in Fig. 3 to generate a voicing decision. Prevoicing is found to be a sufficient, yet not necessary, condition for voicing. Detecting prevoicing is performed by measuring the ratio between the low-frequency mean-rate energy (up to 450 Hz) in the last 20 ms of the closure interval and its maximum value through the whole utterance. If this ratio exceeds a certain threshold, (obtained statistically using histogram analysis), the stop is considered prevoiced. Durations, (i.e., the VOT and the closure duration), are measured using the boundaries generated by the segmentation and categorization system [5] to mark the various segments of the stop consonant.

As shown in Fig. 3, prevoicing is used as the only voicing detection feature for stops that are followed by silences or fricatives (as detected by the segmentation block). For the rest of the stops, it is used as a sufficient condition for voicing. On the other hand, the VOT is usually larger for voiceless stops relative to voiced ones. Histogram analysis showed that two threshold values are needed for accurate voicing detection using the VOT. The choice of the threshold depends on the closure duration as shown in Fig. 3. All thresholds used for the closure duration and VOT are statistically optimized, using histogram analysis and information transmission analysis, to minimize the probability of error during the design phase.

Using the above algorithm for voicing detection yielded an accuracy of 96% as shown in the confusion matrix of Table I. A remark that is worth noting is the interesting role played by the closure duration. Though it does not play a direct role in detecting voicing (i.e., voiced stops did not show systematic closure duration variation relative to unvoiced stops), its indirect role is significant. Attempting voicing detection without the closure duration caused a drop in accuracy from 96% to 90%.

B. Place of Articulation Detection

The first step in the place of articulation detection is to extract the flaps. The flap /dx/ is an allophone of /t/ and /d/ that is used in some dialects in certain contexts (like "matter," "better," etc.). Flaps are characterized by a very short drop in the total energy between two sonorants, which is followed by no release burst and has phonation in it. The duration of the flaps has to be less than or equal to 32 ms. Using these criteria, flaps were recognized correctly with an accuracy of 94%.

The following features are in the place detection of the remaining stops (/t, k, p, d, g, b/):

1) burst frequency (BF);
2) second formant (F2) of the following vowel (VF2);
3) maximum normalized spectral slope (MNSS);
4) burst frequency prominence (DRHF and LINP);
5) formant transitions before and after the stop;
6) voicing decision (using the previous section algorithm).

The burst frequency was statistically found (using information transmission and statistical discriminant analyzes [5]) to be the most important feature for the place detection from the information content standpoint. It is defined as the most prominent peak in the synchrony output during the stop release. The synchrony output is used, as opposed to the mean-rate output, for peak extraction because of its superior ability to extract formants and dominant peaks accurately and its lower sensitivity to noise. The BF for the whole release was taken to be the minimum frequency of the previously mentioned peaks along the whole release duration. This is defined as follows:

$$BF = \min_{j:time\ during\ burst} k_j$$
$$\text{where: } k_j = \arg \max_{i:all\ filters} (ALSD\_output_{ij}) \text{ and:}$$
$$ALSD\_output_{kj} = \max_{i:all\ filters} (ALSD\_output_{ij}). \quad (1)$$

It was found, however, that the burst frequency is highly context dependent. This variability can be significantly reduced by taking the next vowel height into consideration. This *relational invariance* is employed using the second formant location of the following vowel/semivowel (VF2) at the vowel onset point. The points of taking the measurements are determined using the segmentation and categorization program [5]. If there is no following vowel or if the second formant is not clear enough to be extracted, a value of zero is assigned to VF2. Using these two features (BF and VF2), a preliminary place detection is performed using the regions shown in Fig. 4. These regions were designed by the help of unsupervised clustering and Bayesian decision algorithms which showed clear clusters (especially for

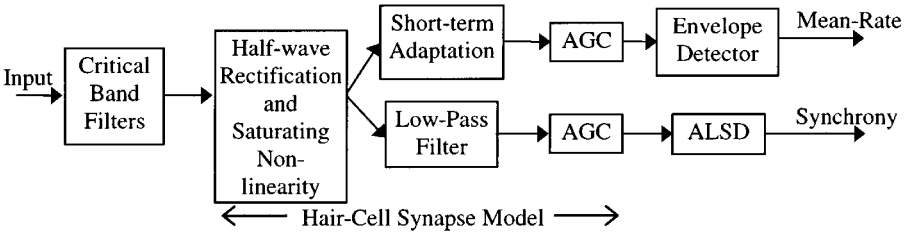

Fig. 2. Block diagram of an auditory-based front-end system.

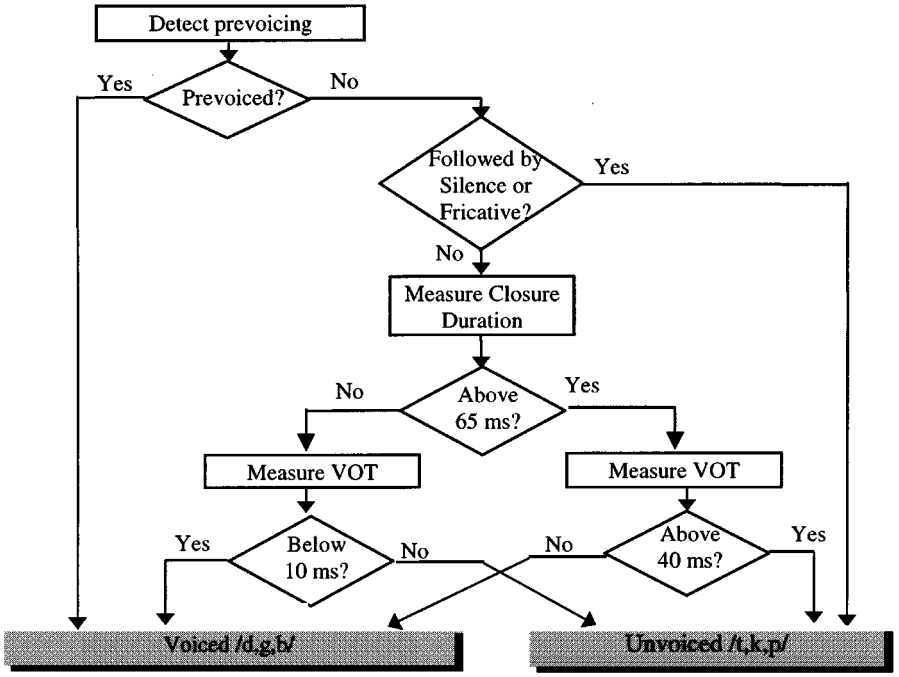

Fig. 3. Algorithm for voicing detection of stop consonants.

TABLE I
CONFUSION MATRIX FOR VOICING DETECTION ON 1200 STOPS. ACCURACY IS 96%

	Detected as voiced	Detected as unvoiced
Voiced	95%	5%
Unvoiced	3%	97%

alveolars and velars) of different places of articulation in the shown regions.

The results of using the BF alone in the place detection are shown in Table II, while Table III shows the result of using both BF and VF2 as shown in Fig. 4. The results are for 270 stops spoken by six different speakers from the TIMIT database. A significant improvement of 10% in the accuracy is clear. This indicates the importance of the vowel context and verifies the concept of relational invariance in the place recognition.

Accounting for the context dependence using Fig. 4 in the classification process also helps normalize for speaker variability. Variations due to speaker gender or dialect are expected to affect the neighboring vowel besides affecting the stop consonant itself. Therefore, the relation between BF and VF2 is less speaker-dependent than BF alone, and hence yields better multispeaker classification results.

It is obvious from Tables II and III and Fig. 4 that the labials are the most missed class when using the burst frequency and the vowel formant. This is in agreement with previous researchers who noted the absence of a prominent peak in labials [23], [45]. They are characterized by a "flat" and weak release spectrum, which is due to the absence of any resonant cavities in their articulation. To improve the detection of labials, the properties of flatness and weakness of their release spectra need to be extracted. The authors developed a new feature called the maximum normalized spectral slope (MNSS). This feature was used in the fricative detection and it proved to be very useful in detecting the dentals [1], [4]. It is defined as

$$MNSS = \frac{\max\limits_{j:release\ burst}\left(\max\limits_{i:all\ filters} Diff(yenv_{ij})\right)}{\max\limits_{j:all\ utterance}\left(\sum\limits_{i:all\ filters} yenv_{ij}\right)} \quad (2)$$

where $yenv_{ij}$ is the ith filter mean-rate (envelope) output at the jth instant, while $Diff()$ is a difference function which approx-

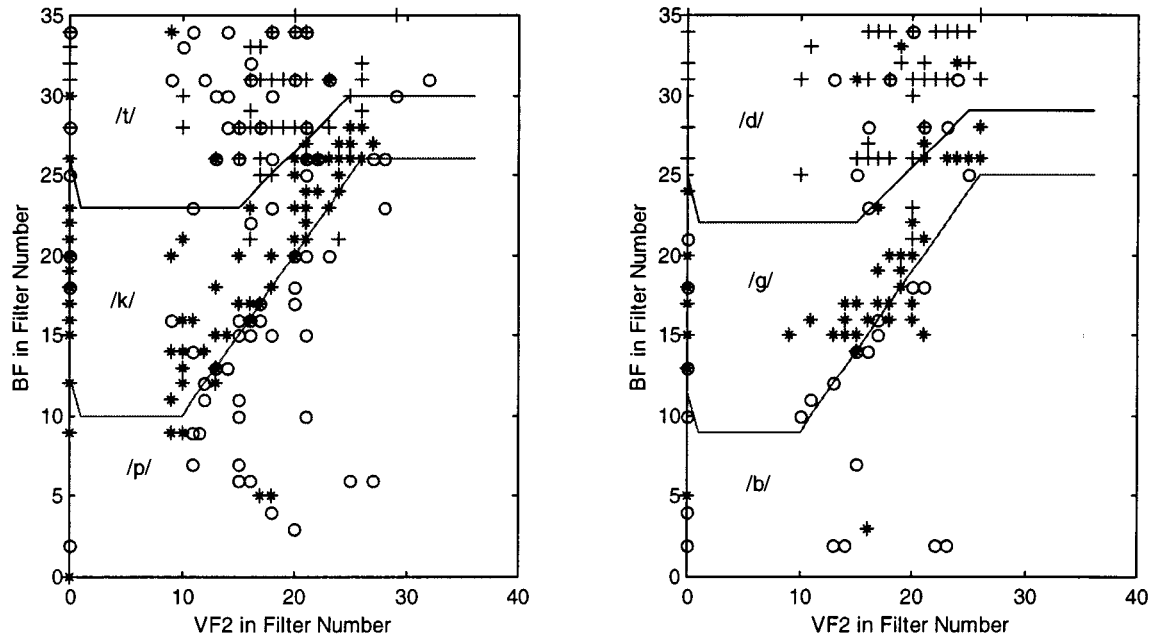

Fig. 4. Two-dimensional space preliminary classification regions for (a) unvoiced stops and (b) voiced stops. Zero VF2 corresponds to the absence of a following vowel's second formant. Alvelars (+), velars (*), and labials (o). It is clear that alveolars and velars show better clustering labials.

TABLE II
CONFUSION MATRIX FOR PRELIMINARY STOP PLACE DETECTION USING THE BURST FREQUENCY (BF) ALONE. TOTAL NUMBER OF STOPS IS 270 FROM SIX DIFFERENT SPEAKERS. ACCURACY IS 72%

	Detected as alveolar /t,d/	Detected as velar /k,g/	Detected as labial /p,b/
Alveolar	74.5%	23.5%	2%
Velar	12.3%	81.5%	6.2%
Labial	14%	28%	58%

TABLE III
CONFUSION MATRIX FOR PRELIMINARY STOP PLACE DETECTION USING THE BURST FREQUENCY (BF) AND VOWEL SECOND FORMANT (VF2). TOTAL NUMBER OF STOPS IS 270 FROM SIX DIFFERENT SPEAKERS. ACCURACY IS 82%

	Detected as alveolar /t,d/	Detected as velar /k,g/	Detected as labial /p,b/
Alveolar	89%	9%	2%
Velar	6.2%	87.6%	6.2%
Labial	14%	20%	66%

imates the derivative with respect to frequency. It could be as simple as the difference between two neighboring filters, i.e.,

$$Diff(yenv_{ij}) = yenv_{ij} - yenv_{(i-1)j}. \tag{3}$$

It is found that a low value of MNSS is a sufficient, but not necessary, condition for labials. The threshold was statistically found (using histogram analysis) to depend on the voicing status of the stop and to be close to the threshold value used in the fricatives, which indicate that this indeed is a characteristic of the labial place of articulation. Stops followed by silences or fricatives, however, do not follow this rule. Those are detected by the segmentation block and the MNSS is not be used with them.

Another aspect of the burst spectrum is the burst frequency prominence. This feature is helpful in discriminating between velars and alveolars. Based on our experiments and comparative analysis of numerous features, two features were developed to describe this property. They are a) the difference between the most dominant peak (MDP) and the energy of the three highest filters, i.e., dominance relative to the highest filters (DRHF) and b) the MDP laterally inhibited by the ten filters above it, call it LINP. The DRHF is defined as

$$DRHF = \max_{j:time\ during\ burst} \left(\max_{i:all\ filters} (ALSD\_output_{ij}) - \max_{k:highest\ 3\ filters} ALSD\_output_{kj} \right). \tag{4}$$

Alveolars are usually characterized by a low value of DRHF due to their high frequency content, and the proximity of their MDP to the highest filters. Therefore, a small DRHF is a necessary condition for an alveolar. The other parameter, LINP, is defined

as

$$
\begin{aligned}
&LINP \\
&= \max_{j:time\ during\ burst} \left(\begin{array}{l} 10 \times ALSD\_output_{kj} \\ - \underset{i=k+1\ to\ k+10}{sum} (ALSD\_output_{ij}) \end{array} \right) \\
&\text{where: } k_j = \arg \max_{i:all\ filters} (ALSD\_output_{ij}) \\
&\text{and: } ALSD\_output_{kj} = \max_{i:all\ filters} (ALSD\_output_{ij}) \\
&\text{and for } k \le i_{\max}; \qquad \text{if } k > i_{\max} \text{ then } LINP = 0 \\
&\text{where: } i_{\max} = 26. \qquad (5)
\end{aligned}
$$

This parameter is used to detect the prominence of the BF peak compared to the filters above it. Large values of LINP were found to indicate a velar, small values of LINP indicate a nonvelar, and moderate values of LINP are ambiguous.

The last feature needed for the place of articulation detection is the formant transitions before and after the stop. These transitions are only applicable if the stop is preceded or followed by a sonorant, as detected by the segmentation and categorization program. Their role was found to depend on whether there is a release or not. For released stops (i.e., stops with a release (burst) segment that is evident in the spectrogram), the formant transitions play only an auxiliary role, while in unreleased stops, their role is primary. In the case of released stops, only *salient* transitions are considered. For a transition to be salient, it has to be of significant slope that exceeds a certain threshold and continuous without sudden jumps or anomalies. Three cases are considered.

1) Clear F2 upward transition to the following sonorant or downward transition from the preceding sonorant. Then, the stop is accepted as labial regardless of the other features.
2) Clear F2 downward transition to the following sonorant or upward transition from the preceding sonorant. Then the stop is accepted as nonlabial regardless of the other features. It is decided whether it is alveolar or velar based on the other features, namely the BF, VF2, DRHF, and LINP.
3) F2 and F3 move away from each other to the following sonorant, or toward each other from the preceding sonorant (velar pinch). In this case, the stop is detected as velar, regardless of the other features involved.

For unreleased stops, the formant transitions (usually preceding the stop) are the only available place cue. Some detailed context-dependent rules were developed to handle those stops [5]. It was not possible, however, to test the correctness of those detailed rules reliably due to the relatively small number of unreleased stops in the database, most of which tend to have clear and strong transitions as explained before and hence do not need the detailed, sonorant-dependent rules. Nevertheless, in the cases tested, the algorithm achieved good accuracy as will be explained later.

This approach is in-line with Dorman *et al.* [20] who found that the significance of the transitions in the human perception process was dependent on their clarity and slope. It also has a practical advantage. Formant transitions are very difficult to measure accurately. Therefore, restricting their use to cases where they are clear, salient and accurately measurable, leads to an improvement in the place detection.

An algorithm, developed to detect the place of articulation using the features and techniques detailed above, is shown in Fig. 5. It gave an accuracy of 90% as shown in Table IV. Performing the same experiment without using the formant transitions causes a 4% drop in accuracy from 90% to 86%. Combining the voicing detection and the place of articulation detection into one system, we obtain a stop classification system. The overall classification accuracy is 86% as shown in Table V.

IV. Discussion

In this work, we developed a new feature-based stop classification system using an auditory-based front-end. The feature-extraction system makes use of both the synchrony and mean-rate outputs. It was clear from our results that the method used in translating the acoustic abstract feature into a measurable parameter has a clear impact on the overall performance. The synchrony is preferred in format/peak extraction (such as the BF), while the mean-rate is used for spectral shapes and amplitudes (such as the MNSS). A new synchrony detector (ALSD) is used to enhance the formant and peak extraction ability. Its ability to detect periodicity and extract dominant peaks accurately is superior to that of the mean-rate envelope detector (an improvement of 5%), and to other synchrony detectors [5]. Repeating the above experiments using the GSD (instead of the ALSD) showed a consistent deterioration of 3% in the place detection on clean and noisy speech. This is attributed to the ALSD's ability to robustly extract the formants while suppressing the spurious peaks [3].

Various acoustic-phonetic features are evaluated for their information content individually and in combination with other features. Some new features were also proposed to describe various aspects of the release spectrum, such as the burst frequency, the spectral flatness, amplitude, compactness, etc. New knowledge-based algorithms were developed to combine the chosen features in the decision making process. These algorithms are designed using a relatively small database (ten speakers) and tested on a much larger database (60 speakers) that was not used in the design process, which demonstrates good generalization ability. They are similar to decision trees, but describe complex interactions between the various features that may be multiple dimensional at some nodes and may depend on the salience of the feature at other nodes. Unlike data-driven approaches, these statistically guided knowledge-based algorithms help improve our understanding of the acoustic–phonetic characteristics of the stop consonants and the complex relation and interaction among various features.

To put the obtained results in perspective, we had to compare them with data-driven systems that rely on huge training databases. Since the databases used in the experiments are different, caution should be exercised when interpreting these comparisons especially when the difference in accuracy is small.

Searle *et al.* [34], in one of the most successful stop consonant recognition experiments, used an auditory-based filter bank and statistical discriminant analysis to detect the place of articulation. They obtained an accuracy of 77% on 148

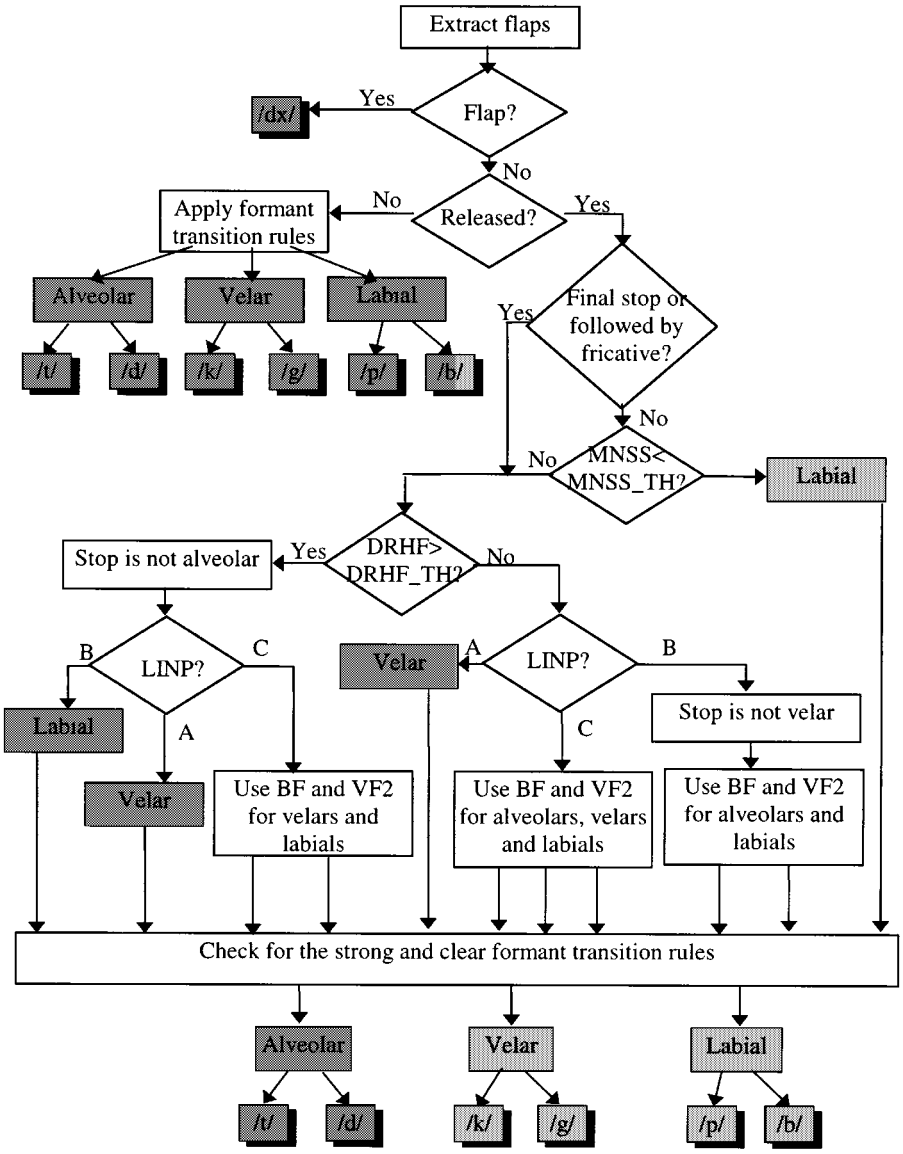

Fig. 5. Hard-decision algorithm for the place of articulation detection of stops. Condition A in the figure is (LINP>LINP_THHI), condition B is (LINP<LINP_THLO), and condition C is [NOT (A OR B)]. MNSS_TH, DRHF_TH, LINP_THHI, and LINP_THLO are the threshold values.

TABLE IV
CONFUSION MATRIX FOR THE PLACE OF ARTICULATION DETECTION ON 1200 STOPS. ACCURACY IS 90%

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

stops. In our experiments, we obtained an accuracy of 90% on 1200 stops. Bush *et al.* [12] obtained classification results ranging between 72% and 81% on 216 stops in syllable initial positions for three male and three female speakers. The results obtained in our experiments show a clear improvement for a much larger database using continuous speech in various syllable positions.

De Mori and Flammia [16] performed phoneme recognition experiments on stops and nasals using back propagation neural networks as classifiers. The stop classification performance was about 82%. This is comparable to the 86% we obtained using a knowledge-based approach.

Nathan and Silverman [29] used time-varying features in a statistical framework to perform place of articulation detection. Their results ranged between 72.3% to 89.1%. On the other hand, Rangoussi and Delopoulos [32] obtained results ranging between 90% and 94% for the place of articulation detection on a smaller testing data set using time-frequency analysis and the

TABLE V
CONFUSION MATRIX FOR THE CLASSIFICATION OF 1200 STOPS. OVERALL ACCURACY IS 86%

	Detected as /t/	Detected as /d/	Detected as /k/	Detected as /g/	Detected as /p/	Detected as /b/	Detected as /dx/
/t/	87.5%	3.5%	5%	0.5%	3%	0.5%	X
/d/	3%	88%	0.5%	7%	0.5%	1%	X
/k/	2.5%	0.5%	87.5%	1%	8%	0.5%	0%
/g/	2%	2%	10%	76%	0	10%	0%
/p/	7%	0%	7%	0%	83.5%	1%	1.5%
/b/	0%	5%	0%	5%	2.5%	85.5%	2%
/dx/	X	X	0%	2%	0%	4%	94%

LVQ classifier. Both results are comparable to the 90% obtained in our work.

Samuelian [33] performed phoneme-level recognition of stops, nasals and liquids using decision trees. He obtained an 83%-90% accuracy for recognition of stops on three speakers. This is comparable to the 86% obtained in this work on a larger number of speakers (60 from seven different dialects). He used statistical tools (namely the C4.5 inductive inference algorithm) to build a decision-tree system. His system however suffered from the inherent traditional limitations of the decision tree algorithms, especially their limited ability to capture multidimensional complex interactions among features like the ones described previously in the place detection algorithm. Moreover, his frame-level recognition did not use the context information as was performed in this work.

V. CONCLUSION

In this work, we investigated the acoustic–phonetic feature-based classification of stop consonants in speaker-independent continuous speech. We used a new auditory-based front-end processing system to generate a dual mean-rate and synchrony representation that combines the advantages of both outputs. Based on the previous research and our own statistical analysis and spectrogram reading experiments, we created a new set of static and dynamic features that are rich in their information content and useful in specific classification tasks. New knowledge-based algorithms were developed to extract the articulatory gestures from these features. Classification experiments were performed on stop consonants extracted from the continuous speech of 60 speakers from seven different dialects of American English in the TIMIT database. The results yielded a 96% and 90% for the voicing and place of articulation detection, respectively. The overall stop classification had an accuracy of 86%. These results demonstrate the importance of using multiple interacting features, context dependence, and relational invariance of features (as opposed to absolute invariance), and emphasize the significance of developing new parameters and algorithms to account for speech variability.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and insightful suggestions.

REFERENCES

[1] A. M. A. Ali, *et al.*, "An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants," in *Proc. IEEE ICASSP'98*, vol. 2, 1998, pp. 961–964.
[2] A. M. A. Ali, *et al.*, "Automatic detection and classification of stop consonants using an acoustic-phonetic feature-based system," in *Int. Congr. Phonetic Sci.*, 1999.
[3] A. M. A. Ali, *et al.*, "Auditory-based speech processing based on the average localized synchrony detection," in *Proc. IEEE ICASSP'2000*, vol. 3, 2000, pp. 1623–1626.
[4] A. M. A. Ali, *et al.*, "Acoustic–phonetic features for the automatic classification of fricatives," *J. Acoust. Soc. Amer.*, vol. 109, pp. 2217–2235, May 2001.
[5] A. M. A. Ali, "Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition," Ph.D. dissertation, Dept. Elect. Eng., Univ. Pennsylvania, Philadelphia, 1999.
[6] G. D. Allen and J. A. Norwood, "Cues for intervocalic /t/ and /d/ in children and adults," *J. Acoust. Soc. Amer.*, vol. 84, no. 3, pp. 868–875, Sept. 1988.
[7] T. R. Anderson, "Speaker independent phoneme recognition with an auditory model and a neural network: A comparison with traditional techniques," in *Proc. ICASSP*, 1991, pp. 149–152.
[8] S. E. Blumstein and K. N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," *J. Acoust. Soc. Amer.*, vol. 66, no. 4, pp. 1001–1017, Oct. 1979.
[9] ——, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Amer.*, vol. 67, no. 2, pp. 648–662, Feb. 1980.
[10] S. E. Blumstien, E. Isaacs, and J. Mertus, "The role of the gross spectral shape as a perceptual cue to place of articulation," *J. Acoust. Soc. Amer.*, vol. 72, pp. 43–50, 1982.
[11] A. Bonneau *et al.*, "Perception of the place of articulation of French stop bursts," *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 555–564, 1996.
[12] M. A. Bush, G. E. Kopec, and V. W. Zue, "Selecting acoustic features for stop consonant identification," in *Proc. ICASSP*, 1983.
[13] R. A. Cole and B. Scott, "The phantom in the phoneme: Invariant cues for stop consonants," *Percept. Psychophys.*, vol. 15, pp. 101–107, 1974 .
[14] T. H. Crystal and A. S. House, "The duration of American–English stop consonants: An overview," *J. Phonetics*, vol. 16, pp. 285–294, 1988 (c).
[15] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 357–366, 1980.
[16] R. De Mori and G. Flammia, "Speaker-independent consonant classification in continuous speech with distinctive features and neural networks," *J. Acoust. Soc. Amer.*, vol. 94, no. 6, pp. 3091–3103, 1993.
[17] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: IV. Sounds with consonants-like dynamic characteristics," *J. Acoust. Soc. Amer.*, vol. 75, no. 3, pp. 897–907, 1984.
[18] B. Delgutte, "Representation of speech-like sounds in the discharge patterns of auditory nerve fibers," *J. Acoust. Soc. Amer.*, vol. 68, pp. 843–857, 1980.
[19] M. F. Dorman and P. C. Loizou, "Relative spectral change and formant transitions as cues to labial and alveolar place of articulation," *J. Acoust. Soc. Amer.*, vol. 100, no. 6, pp. 3825–3830, 1996.
[20] M. F. Dorman *et al.*, "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," *Percept. Psychophys.*, vol. 22, no. 2, pp. 109–122, 1977.

[21] T. J. Edwards, "Multiple features analysis of intervocalic English plosives," *J. Acoust. Soc. Amer.*, vol. 69, no. 2, pp. 535–547, 1981.
[22] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.
[23] ——, *Speech Sounds and Features*. Cambridge, MA: MIT Press, 1973.
[24] M. J. Hunt and Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. ICASSP*, 1989, pp. 262–265.
[25] C. R. Jankowski, H. H. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 286–293, July 1995.
[26] D. Kewley-Port *et al.*, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *J. Acoust. Soc. Amer.*, vol. 73, no. 5, pp. 1779–1793, 1983.
[27] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 73, no. 1, pp. 322–335, 1983.
[28] A. Lahiri *et al.*, "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," *J. Acoust. Soc. Amer.*, vol. 76, no. 2, pp. 391–404, Aug. 1984.
[29] K. S. Nathan and H. F. Silverman, "Time-varying feature selection and classification of unvoiced stop consonants," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 395–405, July 1994.
[30] R. N. Ohde and K. N. Stevens, "Effect of burst amplitude on the perception of stop consonant place of articulation," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 706–714, 1983.
[31] R. K. Potter *et al.*, *Visible Speech*. New York: Van Nostrand, 1947.
[32] M. Rangoussi and A. Delopoulos, "Recognition of unvoiced stops from their time-frequency representation," in *Proc. ICASSP*, 1995, pp. 792–795.
[33] A. Samuelian, "Frame-level phoneme classification using inductive inference," *Comput. Speech Lang.*, vol. 11, pp. 161–186, 1997.
[34] C. J. Searle *et al.*, "Stop consonant discrimination based on human audition," *J. Acoust. Soc. Amer.*, vol. 65, no. 3, pp. 799–809, Mar. 1979.
[35] S. Sandhu and O. Ghitza, "A comparative study of Mel Cepstra and EIH for phone classification under adverse conditions," in *Proc. ICASSP*, 1995, pp. 409–412.
[36] S. Seneff, "A joint synchrony/mean rate model of auditory speech processing," *J. Phonetics*, vol. 16, pp. 55–76, 1988.
[37] ——, "Pitch and spectral analysis of speech based on an auditory synchrony model," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 1985.
[38] R. M. Stern *et al.*, "Multiple approaches to robust speech recognition," in *Proc. DARPA Speech Natural Language Workshop*, Harriman, NY, 1992, pp. 274–279.
[39] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 64, no. 5, pp. 1358–1368, Nov. 1978.
[40] K. N. Stevens and D. H. Klatt, "Role of formant transitions in the voiced–voiceless distinction for stops," *J. Acoust. Soc. Amer.*, vol. 55, pp. 653–659, 1974.
[41] H. M. Sussman *et al.*, "An investigation of locus equations as a source of relational invariance for stop place categorization," *J. Acoust. Soc. Amer.*, vol. 90, no. 3, pp. 1309–1325, 1991.
[42] V. C. Tartter *et al.*, "Perception of intervocalic stop consonants: The contributions of closure duration and formant transitions," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 715–725, 1983.
[43] N. Umeda, "Consonant duration in American English," *J. Acoust. Soc. Amer.*, vol. 61, no. 3, pp. 846–858, 1977.
[44] A. C. Walley and T. D. Carrell, "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 73, pp. 1011–1022, 1983.
[45] V. W. Zue, "Acoustic characteristics of stop consonants: A controlled study," D.Sc. dissertation, Mass. Inst. Technol., Cambridge, 1979.
[46] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Fequenzgrupper)," *J. Acoust. Soc. Amer.*, vol. 33, p. 248, 1961.

Ahmed M. Abdelatty Ali (S'91–M'00) received the B.Sc. and M.Sc. degrees (with distinction and honors) in electrical engineering from Ain Shams University, Cairo, Egypt, in 1991 and 1994, respectively. He received his Ph.D. degree in electrical engineering from the University of Pennsylvania, Philadelphia, in 1999.

From 1991 to 1994, he was a Teaching Assistant with the Electronics and Communication Department, Ain Shams University. He is currently with Texas Instruments R&D, Warren, NJ, and an Adjunct Assistant Professor with the University of Pennsylvania. His research interests include mixed-signal IC design, digital signal processing, and speech processing and recognition.

Dr. Ali is the recipient of the S. J. Stein award for his doctoral research achievements.

Jan Van der Spiegel (M'72–SM'90) received the engineering degree in electromechanical engineering and the Ph.D. degree in electrical engineering from the University of Leuven, Leuven, Belgium, in 1974 and 1979, respectively.

From 1980 to 1981, he was a Postdoctoral Fellow with the University of Pennsylvania, Philadelphia, after which he became an Assistant Professor of Electrical Engineering. In 1987, he became an Associate Professor, and in 1995, Full Professor of electrical engineering. He is currently the Chairman of the Department and Director of the Center for Sensor Technology. His research interests are in analog and digital integrated circuits for intelligent sensors, data acquisition, sensory data processing systems, and acoustic–phonetic feature extraction for automatic speech recognition. He is the Editor for N&S America for *Sensors and Actuators*, and on the editorial boards of the *International Journal of High Speed Electronics* and the *Journal of the Brazilian Microelectronics Society*.

Dr. Van der Spiegel holds the UPS Distinguished Education Term Chair, and was the recipient of the Bicentenneial Chair of the Class of 1940, the Presidential Young Investigator Award and the S. R. Warren and C.&M. Lindback Awards for Distinguished Teaching. He has served on several IEEE Program Committees, and is currently on the Program and Executive Committees of the ISSCC. He is a member of Phi Beta Delta and Tau Beta Pi.

Paul Mueller received the M.D. degree from Bonn University, Bonn, Germany.

He was with the Rockefeller University and the University of Pennsylvania, Philadelphia, and is currently Chairman of Corticon, Inc., Philadelphia. Since 1953, he has worked in molecular and systems neuroscience and has been involved in theoretical studies and hardware implementation of neural networks since the early 1960s.