AN ACOUSTIC-PHONETIC FEATURE-BASED SYSTEM FOR AUTOMATIC PHONEME RECOGNITION IN CONTINUOUS SPEECH

Ahmed M. Abdelatty Ali[(1)], Jan Van der Spiegel[(1)], Paul Mueller[(2)], Gavin Haentjens[(3)] and Jeffrey Berman[(1)]

[(1)]Dept. of Electrical Engineering, University of Pennsylvania, 200 south 33rd St., Philadelphia, PA 19104-6390, USA, [(2)]Corticon, Inc., 155 Hughes Rd., King of Prussia, PA 19406, USA, and [(3)]Dept. of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213-3890, USA.

ABSTRACT

An acoustic-phonetic feature- and knowledge-based system for the automatic segmentation, broad categorization and fine phoneme recognition of continuous speech is described. The system uses an auditory-based front-end processing and incorporates new knowledge-based algorithms to automatically segments the speech into phoneme-like segments that are further categorized into 4 main categories: sonorants, stops, fricatives and silences. The final outputs from the system are 19 class phonemes which contain 7 stops, 6 fricatives, nasals and semivowels, 4 vowel classes and silences. The system was tested on continuous speech from 30 speakers having 7 different dialects from the TIMIT database which were not used in the design process. The results are 92% accuracy for the segmentation and categorization, 86% for the stop classification, 90% for the fricative classification, 75% for the nasal and semivowel extraction and 82% for the vowel recognition. These results compare favorably with previous phoneme classification results.

1. INTRODUCTION

Automatic speech recognition (ASR) has been intensively researched for more than four decades. In the last decade, significant improvements and successes were achieved. However, the understanding of the acoustic-phonetic characteristics of speech, speech variability and speech perception is far from complete. All of the state-of-the-art systems are statistical data-driven systems that rely on training to model our ignorance about speech. Knowledge-based approaches and integrating acoustic-phonetic knowledge in ASR systems has always been an interesting, though elusive, area of research.

Our work is concerned with the design and implementation of a new purely knowledge-based system to segment, categorize and recognize phonemes in continuous speech. The acoustic features that exist in the literature are evaluated and new features are proposed. Both hard-decision and soft-decision algorithms were devised for recognition and to form articulatory-based features such as voicing, manner of articulation and place of articulation. We concentrated more on the obstruent class of phonemes (i.e. fricatives, stops and affricates) due to their noisy, dynamic, relatively short, weak, speaker- and context-dependent nature which made them one of the most challenging phonemes to automatically recognize in continuous speaker-independent speech.

This system could be used directly for knowledge-based phoneme recognition, or it could be used as a front-end for a statistical ASR system (such as Hidden Markov Model (HMM) or Artificial Neural Network (ANN) systems). The soft-decision algorithms of our system generate probability estimates (certainty factors) for each phoneme class that could be used as inputs to an HMM or ANN system. The system was designed with the goal of performing acoustic-phonetic recognition while minimizing error propagation and loss of information in order to enable its integration with statistical ASR systems.

2. SYSTEM DESCRIPTION AND RESULTS

A block diagram of the system is shown in Fig. (1) and an example of the output is shown in Fig. (2). The front-end processing is an auditory-based system that is described in detail in [2,6,13]. The output of the front-end processing is passed to the segmentation and categorization system [3,6] that uses the following features:

1. Total energy.
2. Spectral Center of Gravity (SCG).
3. Duration.
4. Low, medium and high frequency energy.
5. Formant transitions.
6. Silence detection.
7. Voicing detection.
8. Rate of change of energy in various frequency bands.
9. Rate of change of SCG.
10. Most prominent peak frequency.
11. Rate of change of the most prominent peak frequency.
12. Zero-crossing rate.

Using the above features in the rule-based algorithm described in detail in [3,6], stops, fricatives, sonorants and silences were extracted successfully with a 92% accuracy (4% substitution, 3% insertion and 1% deletion), when tested on continuous speech from 30 speakers of 6 different dialects of American English from the TIMIT database (more than 7000 phonemes from 300 sentences).

0-7803-5474-5/99/$10.00(C)1999 IEEE

After the extraction of the fricatives and the stops, a classification system is encountered to classify the different phonemes. The classification of the fricatives is divided into voicing detection and place of articulation detection. The feature used in the voicing detection of fricatives is the duration of the unvoiced portion (DUP) where voicing is measured by the presence of low frequency energy in either the mean-rate or the synchrony outputs. If the DUP is below an empirically determined threshold, the fricative is detected as voiced, otherwise it is unvoiced.

The place of articulation detection of fricatives is performed using the following features:

1. The Most Dominant Peak (MDP) frequency from the synchrony detector.
2. The Maximum Normalized Spectral Slope (MNSS), defined as the ratio of the largest spectral slope of the mean-rate spectrum to the maximum total energy in the utterance. It demonstrates the flatness and weakness properties which characterize labial and dental phonemes.
3. The Spectral Center of Gravity (SCG).
4. The Most Dominant Spectral Slope (MDSS) from the synchrony output.
5. The Dominance Relative to the Highest Filters (DRHF), defined as the difference between the MDP synchrony value and that of the highest 3 filters.

The above features were extracted and manipulated using both hard-decision (binary output) and soft-decision (certainty factors or probability estimates) algorithms. Soft decision algorithms are richer in information and more useful especially for generating features (probability estimates) that could be used by a following classification system while minimizing any information loss due to an erroneous decision.

Stop detection is also divided into voicing detection followed by place of articulation detection. The features used in the voicing detection are:

1. Prevoicing, defined as presence of voicing during the closure period.
2. Voicing Onset Time (VOT), defined as the duration from the release to the start of voicing.
3. Closure duration.

The place of articulation detection of stop consonants was performed using the following features:

1. The Burst Frequency (BF), defined as the frequency of the most prominent peak during the release burst of the stop.
2. The second formant of the following vowel (in prevocalic stops).
3. The Maximum Normalized Spectral Slope (MNSS), similar to that used for the fricatives.
4. The burst frequency prominence, as described by two features, namely the DRHF described above, and the LINP which is the value of the most prominent peak of the synchrony response after being laterally inhibited by the higher 10 filters.
5. Formant transitions before and after the stop.
6. The voicing decision.

3. CONCLUSION AND COMPARISON WITH PREVIOUS WORK

Despite the recent successes in the Automatic Speech Recognition (ASR) field, the acoustic-phonetic characteristics of speech and their variability with context and speaker are not fully understood yet. More research is still needed to achieve a good understanding of this topic in order to build improved front-end processing systems that are able to extract the useful, information-rich, acoustic features. This knowledge is expected to have a profound effect on the automatic speech recognition systems whose performance can significantly improve by integrating more knowledge into their design.

Our work is concerned with this problem. We studied the acoustic-phonetic characteristics of continuous speech from multiple speakers with different dialects from the TIMIT database. The acoustic features described in our work and the algorithms developed to extract them are expected to be a significant contribution to the acoustic characterization and the automatic recognition of continuous speech. Our work builds on the previous work in this area and introduces modifications and additions that resulted in profound improvements in the overall system performance.

This work could be exploited in Hidden Markov Models (HMMs) or Artificial Neural Network (ANN) speech recognition systems. These systems could make use of the designed system outputs in order to improve the front-end processing and create additional inputs to the data-driven (training-based) classifiers. These inputs are rich in information and incorporate considerable speech knowledge in their design. This is especially true with the soft-decision algorithms whose output is in the form of posterior probability estimates which are compatible with HMMs and could be used as additional inputs that are rich in information and independent of speaker or context. They can also be used in a knowledge-based acoustic-phonetic speech recognition system by making a hard decision directly from the probability estimates.

The acoustic-phonetic recognition system developed was tested on the TIMIT database continuous speech of 30 speakers from 6 different dialects of American English. Since no similar system was developed before, we are going to compare the results of the system's *subtasks* with those of previous research. For the segmentation and categorization, an accuracy of 92% is achieved which compares favorably with the best previous results obtained by Liu [11], who used landmark detection for this purpose and obtained 90% on a similar task. The accuracy of extracting the nasals and semivowels and the recognition of the vowels categories were 75% and 82% respectively which are comparable to previous results [9]. For fricatives, a recognition accuracy of 93% is achieved for the place of articulation, 95% for voicing detection and 90% for the fricative classification. The best previous result was 77%-80% for place detection and 83% for voicing detection [10,14]. For the stops, 90% accuracy is achieved for the place detection, 97% for the voicing detection and 86% for the stop classification. These results compare favorably too with the

0-7803-5474-5/99/$10.00(C)1999 IEEE

best previous which was 75%-80% classification accuracy using knowledge-based approaches and 82% using statistical (training-based) approaches [7,8,12]. A detailed comparison is given by Ali in [1-6]. Such significant improvement is mainly due to the use of auditory-based front-end processing, and new feature extraction techniques and manipulation algorithms, which integrate several acoustic properties in the decision making process.

4. ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF REU Grant no. EEC96-19852) and Catalyst Foundation.

5. REFERENCES

[1] Ali, A.M.A., *et al*., "Acoustic-phonetic Features for the Automatic Recognition of stop consonants", *Journal of the Acoustical Society of America*, 103(5), pp. 2777-2778, 1998.

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

[3] Ali, A.M.A., *et al*., "Automatic detection and classification of stop consonants using an acoustic-phonetic feature-based system", *14th International Congress of Phonetic Sciences*, (accepted), 1999.

[4] Ali, A. M. A., "Acoustic features for the automatic recognition of fricatives", Technical Report, TR-CST27AUG97, Center for Sensor Technologies, University of Pennsylvania, 1997.

[5] Ali, A.M.A., "Acoustic-phonetic Features for the Automatic Recognition of Stop Consonants", Technical Report, TR-CST22DEC97, Center for Sensor Technologies, University of Pennsylvania, 1997.

[6] Ali, A. M. A., "Segmentation and Categorization of phonemes in continuous speech", Technical Report, TR-CST25JUL98, Center for Sensor Technologies, University of Pennsylvania, 1998.

[7] Bush, M. A., *et al.*, "Selecting acoustic features for stop consonant identification", *Proc. ICASSP*, 1983.

[8] De Mori, R. and Flammia, G., "Speaker-independent consonant classification in continuous speech with distinctive features and neural networks", *J. Acoust. Soc. Am.*, 94 (6), pp. 3091-3103, 1993.

[9] De Mori, R. and Suen, C.Y., "New Systems and Architectures for Automatic Speech Recognition", Springer Verlag, 1985.

[10] Hughes, G. W. and Halle, M., "Spectral Properties of Fricative Consonants", *J. Acoust. Soc. Am*., 28, pp. 303-310, 1956.

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

[12] Searle, C. J. *et al*, "Stop consonant discrimination based on human audition", *J. Acoust. Soc. Am.*, 65 (3), pp. 799-809, 1979.

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

[14] Stevens, K.N., et al, "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters", *J. Acoust. Soc. Am*., 91, pp. 2979-3000, 1992.

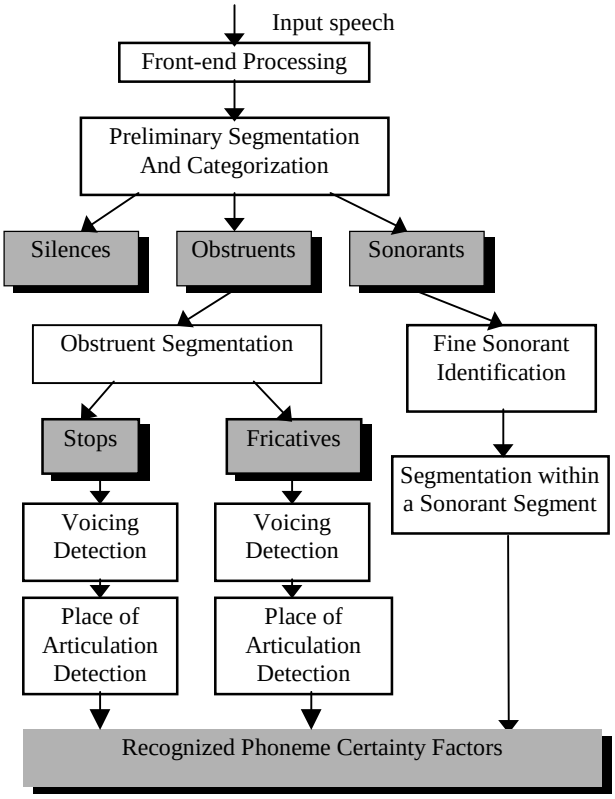

Fig. (1) (a) Block diagram of the system.

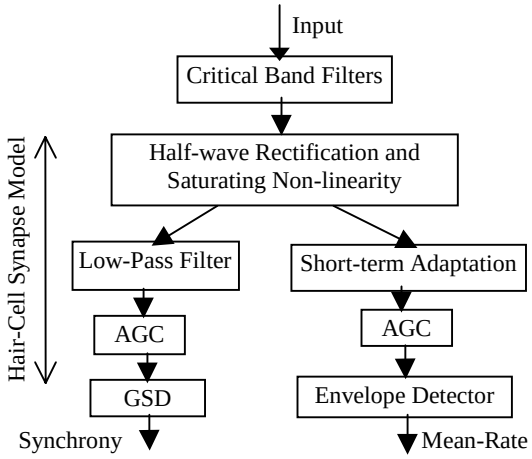

Fig. (1) (b) Block diagram of the auditory-based front-end processing system.

0-7803-5474-5/99/$10.00(C)1999 IEEE

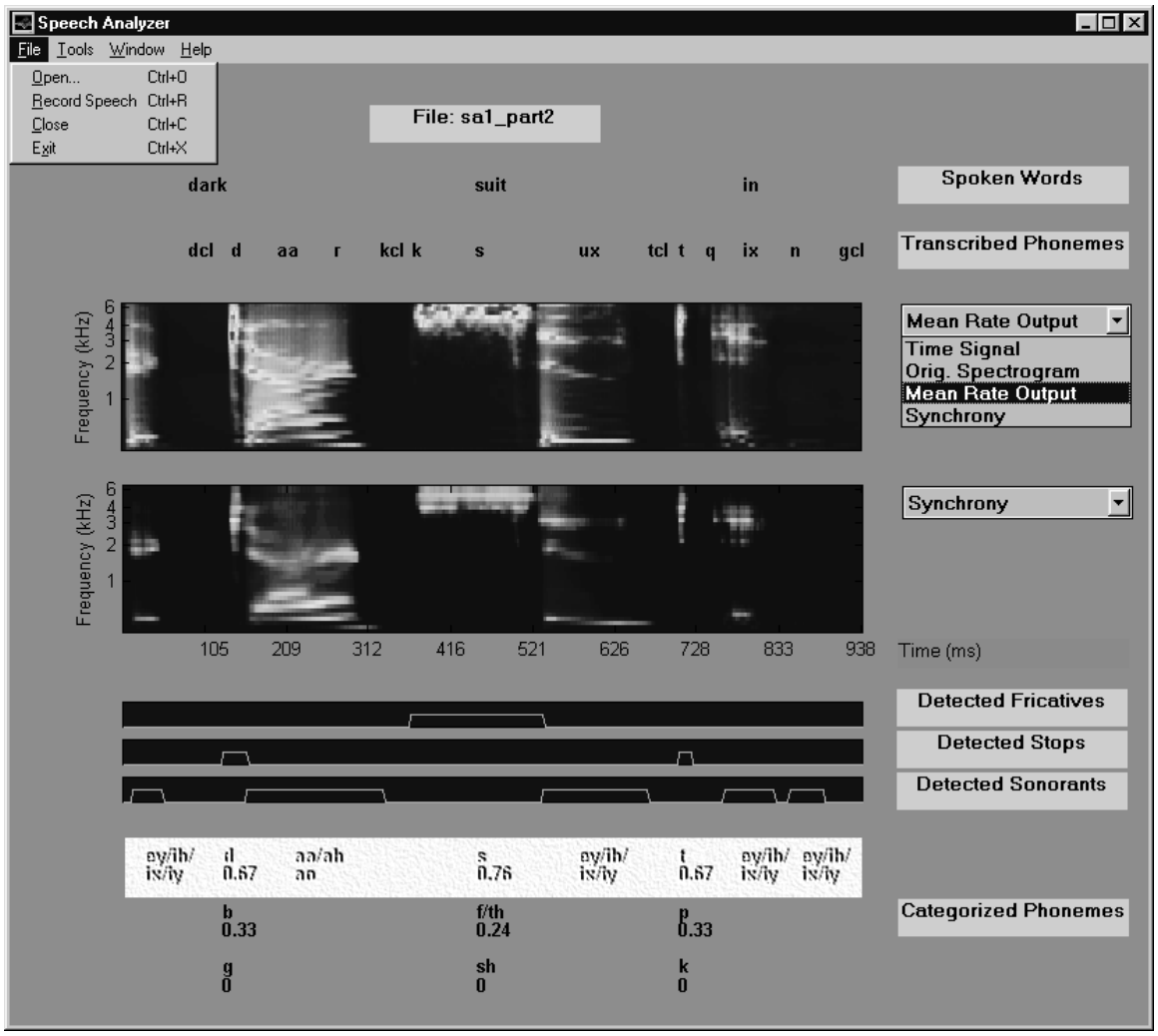

Fig. (2) An example of the output from the system for the phrase “dark suit in” spoken by a female speaker. It shows the two types of pseudo-spectrograms (mean-rate and synchrony), the detected categories (fricatives, stops and sonorants) and the fine phoneme recognition. The recognized phonemes are listed at the bottom of the figure with the corresponding certainty factors.

0-7803-5474-5/99/$10.00(C)1999 IEEE