AUDITORY-BASED SPEECH PROCESSING BASED ON THE AVERAGE LOCALIZED SYNCHRONY DETECTION

Ahmed M. Abdelatty Ali [(1)], *Jan Van der Spiegel* [(1)] *and Paul Mueller*[(2)]

[(1)] Department of Electrical Engineering,
University of Pennsylvania,
Philadelphia, PA 19104-6390

[(2)] Corticon, Inc.
155 Hughes Rd,
King of Prussia, PA 19406

ABSTRACT

In this paper, a new auditory-based speech processing system based on the biologically rooted property of average localized synchrony detection (ALSD) is proposed. The system detects periodicity in the speech signal at Bark-scaled frequencies while reducing the response's spurious peaks and sensitivity to implementation mismatches, and hence presents a consistent and robust representation of the formants. The system is evaluated for its formant extraction ability while reducing spurious peaks. It is compared with other auditory-based front-end processing systems in the task of vowel recognition on clean speech from the TIMIT database and in the presence of noise. The results illustrate the advantage of the ALSD system in extracting the formants and reducing the spurious peaks. They also indicate the superiority of the synchrony measures over the mean-rate in the presence of noise.

1. INTRODUCTION

Due to the superb ability of humans to recognize speech in noisy environments, auditory-based front-end processing systems were developed to emulate some of the processing performed in the human auditory periphery. Several speech recognition experiments have demonstrated that such auditory-based systems yield better performance (in terms of recognition accuracy), in the presence of noise, compared to the traditional LPC and the Mel-Frequency Cepstral Coefficients (MFCC) [5][6]. The relatively robust performance of the auditory-based systems was attributed to the Bark-scale filtering, the compressive non-linearity, the short-term adaptation, the forward masking and the synchrony detection [5][9].

Despite their superior performance, most auditory-based systems are characterized by a relatively long processing time. This makes real-time software implementation of ASR systems, which use such front-ends, difficult with the present computation power. Hardware implementation of the front-end system, using parallel analog processing, is more economic for real-time operation to be achieved.

In this work, we investigate some of the auditory-based systems that proved to yield relatively good and robust performance, and are readily implementable in analog VLSI technology. Those include the Bark-scaled filter bank mean-rate output, the Lateral Inhibitory Network (LIN) output [12], and the Generalized Synchrony Detector (GSD) output [10][11]. A new system is developed by the authors as a modification to the GSD. It is called the Average Localized Synchrony Detector (ALSD) [4] and is designed to alleviate some of the limitations of the GSD. These four systems are compared in their formant extraction ability as indicated by vowel recognition experiments for multiple speakers with 7 different dialects of American English from the TIMIT database using the first two formants.

2. AUDITORY-BASED PROCESSING

The general structure of the auditory-based processing systems tested in this work is shown in Fig. (1). It consists of a Bark-scaled filter bank of 36 filters with a spacing of half a Bark between neighboring filters. The filter-bank used here is a simulation of an actual analog cochlea that was implemented in VLSI [7][8]. This choice of the filter-bank is made in order to ensure its practicality from the hardware implementation standpoint. Besides the critical-band filtering, the system includes other auditory effects like compressive non-linearity, half-wave rectification, automatic gain control, short-term adaptation and forward masking [4]. It gives two outputs: the mean-rate output and the synchrony output. The synchrony detector block could be a LIN, GSD or ALSD depending on which system is being tested.

The LIN system is based on inhibiting the output of each filter by neighboring filters [12]. This could be as simple as subtracting the neighboring filter output, or it could be more involved like using a feedforward or feedback inhibitory network. The approach used in this work is using a feedforward lateral inhibitory network similar to that used by Shamma and described in [12]. The output of each unit is computed by subtracting a weighted sum of its neighbors, followed by a threshold operation and a time-window average. In this way, the peaks (formants) are enhanced by detecting the filters that have strong phase differences with their neighbors. This is a simple, fast and effective approach for detecting the peaks and producing a robust formant representation.

The GSD system is based on the generalized synchrony detector developed by Seneff [10][11]. The GSD computes an auto-correlation-like output by finding the soft-limited ratio of the expected (averaged) value of the sum and difference of the output of each filter and a delayed version of it. The delay of each GSD must match its corresponding filter's center frequency (i.e. the delay is equal to the inverse of the center frequency). The GSD enhances the formants and improves the spectral resolution by detecting the periodicity (temporal structure) in the filter outputs instead of the envelope (mean-rate).

0-7803-6293-4/00/$10.00 ©2000 IEEE.

Despite its advantage over the mean-rate response in enhancing and extracting the formants and its better performance in the presence of noise, the GSD has some limitations. First, it is relatively slow. Second, it suffers significant spurious peaks due to individual harmonics of the fundamental frequency and other artifacts. Those peaks could be so strong in the low frequency filter responses for female speakers that they affect the detection of the first formant [11]. Third, it requires accurate matching between the delay time used in the GSD and the center frequency of the corresponding filter. This matching, which needs to be as tight as 0.1%, is so critical that it necessitates oversampling for the high frequency filters otherwise significant spurious peaks appear [11]. Although such accurate matching is possible in software, it is difficult to achieve in hardware due to the inevitable component and dimension tolerances in VLSI technology. Such a limitation becomes increasingly important for a system like the GSD whose software implementation is too slow for practical applications.

The way to reduce the spurious peaks and relax the sensitivity to matching is to increase the filter bandwidth [11]. This however will significantly deteriorate the resolution and performance of the system in a way that defeats its original purpose. In other words, in the absence of exact matching, there is an accuracy-resolution trade-off that can not be solved. To get rid of spurious peaks that affect the formant extraction accuracy we need to smooth the spectrum by using wider-band filters, which deteriorates the resolution.

To alleviate the previously mentioned problems, we modified the GSD in order to represent the average localized synchrony [13]. The output of each ALSD is the average of n GSD's tuned to the same frequency but applied to several filters in the neighborhood of the filter corresponding to that frequency. The number n depends on the resolution and bandwidth of the filters used. This can be expressed as follows:

$$ALSD_i = \frac{1}{n} \sum_{k=i-n1}^{i+n2} GSD_i(y_k) \qquad (1)$$

where $ALSD_i$ is the ALSD output of the ith channel (filter); GSD_i is the output of the GSD which is tuned to the ith filter; y_k is the output of the kth filter (after the AGC stage); $GSD_i(y_k)$ is the output of the ith GSD (i.e. the GSD tuned to the ith filter) when applied to the kth filter. The constants $n1$ and $n2$ add up to n. (i.e. $n=n1+n2$). We chose n to be equal to 3, with one filter on each side of the center filter. (i.e. $n1=n2=1$ and k ranges from i-1 to i+1). We need to emphasize that the operation mentioned in equation (1) is not equivalent to simply averaging the inputs of neighboring filters and applying them to the same GSD. It is also different from averaging the outputs of neighboring GSD's. The non-linearity of the GSD and its tuning characteristics make the ALSD output significantly different from those two averaging operations [4]. This is shown in Fig. (2).

The ALSD provides an extra degree of freedom, which enables us to achieve smoothing while preserving resolution. It also decreases the system response to individual harmonics (compared to formants) because the harmonics are usually limited to one filter, while formants extend to neighboring filters. By averaging the GSD responses in the vicinity of the filter, responses to individual harmonics will be relatively attenuated.

Detailed description, illustration, and verification of the ALSD operation are found in [4]. The system demonstrated that it was capable of achieving the same resolution as the GSD (as evidenced by its ability to resolve three sinusoidal signals that are spaced approximately one Bark apart) while considerably smoothing the spectrum [4]. An example of the system's response to speech is shown in Fig. (2). The spurious peaks in the GSD, LIN and Mean-Rate responses have been significantly reduced by the ALSD, while preserving the formants. On the other hand, the wider-filter system in Fig. (2e) destroyed the formants while retaining the harmonic that is below F1. The averaged-output system in Fig. (2f) also failed to reduce the harmonic and yielded a distorted formant structure. Therefore, we can see that the ASLD system is capable of achieving *selective* smoothing whereby it reduces the spurious peaks while preserving the formants.

3. EXPERIMENTS AND RESULTS

The four auditory-based systems, namely the mean-rate, the LIN, the GSD and the ALSD, are tested for their formant extraction ability from clean speech and in the presence of noise. The test is in the form of a four-vowel recognition experiment performed on continuous speech from multiple speakers with 7 different dialects of American English from the TIMIT database. The four vowels are: /ae/, /iy/, /aa/ and /uw/. They are chosen to represent the four main tongue positions and hence could be classified by the first two formants. The experiments are performed on clean speech and on speech distorted by additive white Gaussian noise with different signal-to-noise ratios.

The first two formants are extracted using a relatively simple formant tracking algorithm. The algorithm picks the peaks that satisfy certain location-, amplitude- and continuity-constraints. The same algorithm is used for all the systems. The only difference is in the amplitude thresholds, which are optimized independently for each system.

Using the first two formants, the four vowels are classified using two threshold values which divide the two dimensional space into four regions using Bayesian classification with the maximum posterior probability criterion. The system is trained using six speakers (3 males and 3 females) and tested on 30 different speakers with more than 1000 of the above vowels. Three measurements are taken in the middle third of each vowel and a majority rule is employed.

The choice of this classification method is motivated by the purpose of the experiments. We are interested in evaluating the systems' abilities to accurately extract the formants in the presence of noise. Thus it is necessary to ensure that the classification decision is based on the formant positions and not on any other spectral artifacts. Moreover, we need to evaluate the ability of the system to reduce spurious peaks and hence enable us to use a relatively simple formant tracking algorithm.

The results of the experiments are summarized in Fig. (3). For clean speech, we see that the ALSD gives the best performance, followed by the mean-rate, the LIN and finally the GSD. The relatively bad performance of the LIN and the GSD is attributed to the presence of spurious peaks that cause errors in formant extraction. The ALSD smoothes the response, while preserving the formants, and hence improves the performance. When noise is added to the system, the performance deteriorates. The deterioration is worst for the mean-rate, which falls sharply. This is in agreement with previous findings which demonstrated that synchrony measures are usually more robust than the mean-rate. We can also see that the deterioration of the ALSD response with noise is almost identical to that of the GSD. This indicates that the ALSD preserved the robustness of the GSD while improving the performance by decreasing the spurious and individual harmonic peaks. This selective smoothing improves the performance significantly when using simple formant tracking algorithms like the one used in our experiments.

4. CONCLUSION

A new auditory-based speech processing system based on the average localized synchrony detection (ALSD) is developed to alleviate some of the limitations of the GSD, such as the presence of spurious peaks, sensitivity to implementation mismatches, and response to individual harmonics. The system is compared with several other auditory-based systems in their formant extraction ability from clean and noisy speech. The other systems are the Bark-scaled mean-rate, the lateral inhibitory network (LIN) detector, and the generalized synchrony detector (GSD).

The results demonstrate the advantage of the ALSD in extracting the formants and reducing the spurious peaks. They also indicate the superiority of the synchrony measures, in the presence of noise, compared to the mean-rate. In spite of their superb formant extraction ability, the LIN and GSD are plagued by significant spurious peaks, which complicate the formant-tracking task. Such spurious peaks are significantly reduced by the ALSD, which yields a better performance with the relatively simple formant-tracking algorithm used in the experiments.

5. ACKNOWLEDGMENT

This work is supported by a grant from Catalyst Foundation.

6. 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", *Proc. 14th International Congress of Phonetic Sciences*, pp. 1709-1712, 1999.

[4] Ali, A.M.A., "An average localized synchrony detector (ALSD) for an auditory-based front-end speech processing system", Technical Report, TR-CST10AUG99, Center for Sensor Technologies, University of Pennsylvania, 1999.

[5] Ghitza, O., "Auditory models and human performance in tasks related to speech coding and speech recognition", *IEEE Trans. Speech and Audio Proc.*, 2(1), pp. 115-132, 1994.

[6] Jankowski, C.R., Vo, H.H. and Lippmann, R.P., "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 286-293, 1995.

[7] Liu, W., *et al.*, "Voiced-speech representation by an analog silicon model of the auditory periphery", *IEEE Trans. Neural Networks*, 3(3), pp. 477-487, 1992.

[8] Lyon, R. and Mead, C., "An analog electronic cochlea", *IEEE Trans. Acoust., Speech and Signal Processing*, 36(7), pp. 1119-1134, 1988.

[9] Ohshima, Y., "Environmental robustness in speech recognition using physiologically-motivated signal processing", Ph.D. thesis, Carnegie Mellon University, 1993.

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

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

[12] Shamma, S., "The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives", *J. Phonetics*, 16, pp. 77-91, 1988.

[13] Young, E. D. and Sachs, M. B., "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers", *J. Acoust. Soc. Am.*, 66(5), pp. 1381-1403, 1979.

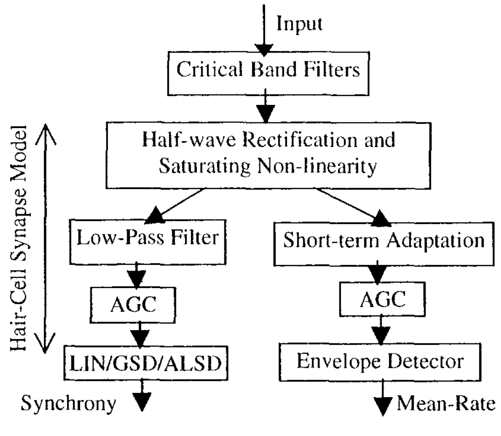

Fig. (1) Block diagram of the auditory-based front-end processing system.

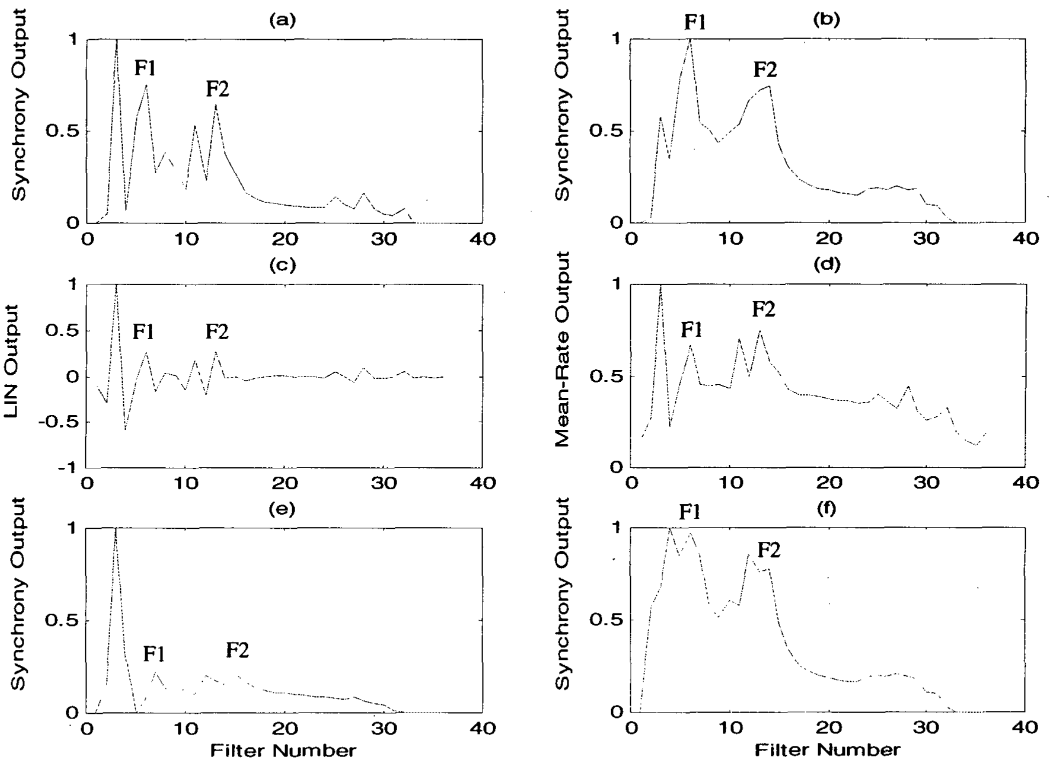

Fig. (2) System response for the vowel /aa/ spoken by a female speaker. (a) GSD, (b) ALSD, (c) LIN, (d) Mean-Rate, (e) GSD with averaged inputs (wider filters), (f) GSD with averaged outputs.

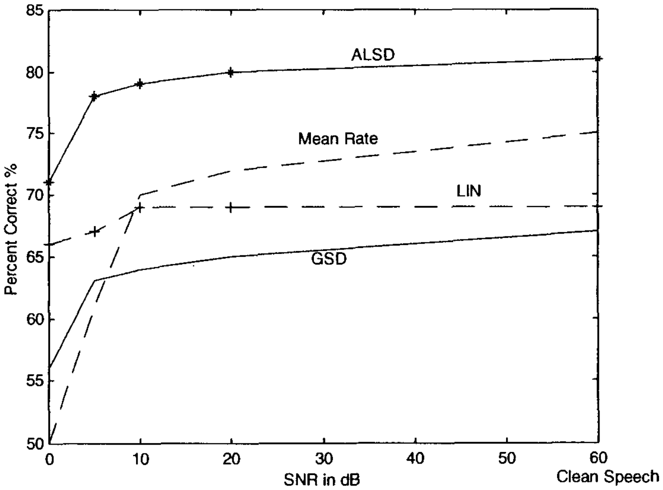

Fig. (3) Classification accuracy for the different systems in vowel recognition experiments on 30 speakers with 7 different dialects of American English from the TIMIT database at various noise levels. The ALSD gives the best performance. The mean-rate output deteriorates more sharply with noise than the synchrony measures.