ROBUST FEATURE SETS FOR THE SPEECH PROCESSING

Zhandos YESSENBAYEV
L.N. Gumilev Eurasian National University, Astana, Kazakhstan
5 Munaitpasov Str., Astana, 010008, +7 (7172) 35-38-06, yessen@mail.ru

Abstract: This paper compares the various feature sets for speech signal such as cepstral-based and acoustically-driven parameters to analyze and see how robust they are in different noise environments. Despite the success of the cepstral-based features in the tasks of speech processing, they are still susceptible to noise. In this work, we explored how to extend MFCC-based feature sets with other acoustically-driven parameters to add some robustness for the segmentation of speech. Experiments conducted on TIMIT dataset using the standard HMM/GMM framework show that better performance can be achieved if cepstral-based features are combined with acoustic parameters.

Keywords: *MFCC, acoustic parameters, HMM, speech segmentation, TIMIT.*

I. INTRODUCTION

One of the important aspects of speech for computer science is speech recognition, which is an attempt to automate the "understanding" of speech by machines. The ability of a computer to "understand" speech and act accordingly would potentially reduce the human-load and the risk in human-dependent applications. The real applications of automatic speech recognition (ASR) systems are widely used in telecommunication, medicine, and military. Although, the success of such systems is notable, they all suffer from noisy environment and the performance degrades significantly. And because of this instability, they hardly can substitute human in the applications where the responsibility is high. Therefore, the robustness of such systems plays very important role. One question that consequently arises is that whether there is a unique representation of speech signal which helps human extract only the information he needs. So our concern is the quest for better (in the sense of robustness) description of speech, which will benefit the current systems and prevent them from failing in the presence of noise, that is, will provide some stability and reliability in the speech recognition process.

Our hypothesis is that the combination of cepstral coefficients and acoustic parameters can be beneficial for the ASR systems in the sense of robustness. On the other hand, we are looking for the reliable and, yet, simple acoustic parameters, which are cheap to extract and, therefore, leaving a computational gap for the later stages. And as a task of speech processing, we focus on the segmentation of a speech signal into two wide classes – sonorant and obstruent, which comprise broad phoneme classes of American English language. The sonorant class consists of vowels, liquid and glides (semivowels), nasals, whereas the obstruent class consists of fricatives, affricates and stops [1]. We also add non-speech parts (e.g. silence) of a speech signal into obstruent class. The accurate realization of this task is crucial for the systems that extract the phonetic constituents from these regions, because the error in this stage will propagate further causing pyramidal effect on error rates. To get competitive performance, we will invoke HMM-based approach using GMM as the observation probability density function, since this approach has proved to be computationally feasible and successful in speech recognition tasks.

The rest of the work is organized as follows. In Section 2 we discuss related work done by other researchers regarding our problem. Section 3 is dedicated to the analysis of data. Here we provide the description and behavior of several acoustic parameters on the input data. Sections 4 and 5 present the results obtained and their careful discussion. And, finally, Section 6 concludes the work.

II. RELATED WORK

There are almost no work done on segmentation of speech signal using such a combination of parameters, although there are applications where cepstral coefficients and APs are used together to solve one common problem but each in the different subtasks, for instance, using the first parameters as a feature set and the second as a threshold for the obtained result. Therefore, we believe that our contribution and impact to the ASR systems carries a significant importance and will be truthfully appreciated.

The problem of sonorant-obstruent segmentation has attracted researchers for its importance in the general hierarchy of tasks in the automated speech recognition systems based on the phonetic transcriptions, syllable and

Manuscript received January 31, 2012; revised May 29, 2012

sub-word detections from the broader regions of interest. Hence, the correctness of the segment boundaries directly affects the performance of such detectors and the mismatches made in the initial stages are difficult to recover later on. An example of such systems is a work done by A. Jansen and P. Niyogi [2], which laid as a basic source of ideas and was a motivation for the current work.

The approach proposed uses the notion of distinctive features and exploits the idea of landmark detectors, which, they claim, could be an alternative to the modern HMM-based approaches. A hierarchical model is based on the number of feature detectors that output a set of candidate landmarks, which are, then, probabilistically integrated to construct the most likely sequence of broad classes. Although the model built is up to the broad-class level, it explicitly includes the sonorant-obstruent segmentation stage. For the segmentation, an SVM was trained on 39-dimensional mel-frequency cepstral coefficient feature sets. The accuracy was estimated with the measure they proposed, which is, basically, the percentage of the phonemes that fall into corresponding sonority regions given the threshold of being "accepted" by those regions (for details, see the section "Experimental Results"). In spite of the high performance for different thresholds, what interested us most is that the difference between the respective sonorant and obstruent measures is high and grows significantly as the threshold is increased. One of the reasons could be that the segmenter assigns wider regions for one class while narrower for the other, what makes such a difference in the measures. Therefore, the question of the quality of the segmentation arises. Another issue that was not considered in this paper is noise.

In [3] a frame-based SVM classification using the general-purpose MFCCs is built, where the problem of noisy condition is addressed. For that, authors estimate signal-to-noise ratio from the frame energy histograms, noticing that, for stationary noise, there is going to be two peaks: one corresponds to the accumulations of non-speech frames containing only noise, and the other – to the speech plus noise portion. The difference between these two peaks gives a measure that is a "good indicator" of SNR. Then, based on this measure they vary adaptively the parameter λ in the classification rule:

$$x_i \in \{sonorants\} \Leftrightarrow w_0^T x_i + b_0 > \lambda. \qquad (1)$$

Unclear part of this work is a map from SNR to optimal threshold during the training stage. Another problem is that the comparison of the results for noisy data was made between different settings of λ, but not with the results obtained on clean data. For example, the difference in performance between clean and pink noise added data goes above 10%, what requires the explanation how "good" or "bad" it is, i.e. the measure of accuracy estimation was not given. In addition, this type of approach doesn't take into account the noisiness of the extracted feature sets.

An alternative to these two papers in terms of data representations is a work by A. Juneja and C. Espy-Wilson [4]. Their method is based on the extraction of different acoustic parameters and passing the relevant parameters to SVMs trained for each broad class. Although it is not very clear how some of the parameters are obtained (third formant of a speaker, or probability of voicing), the idea of separating the parameters according to the broad class should really be appreciated. The attempt to compare the performances of their system with HMM-MFCC-based one fails in that they built not quite a competitive model for the second system, which has the performance not comparable to the state-of-the-art HMM-based systems.

Some other studies using APs as a basis of speech representation are described in [5-7]. All the parameters extracted are quite simple yet natural, however, the robustness to noise should be inspected carefully. For example, Wiener entropy may not reflect structural properties of a signal in the presence of noise as well as the periodicity estimation of noisy and non-stationary signals is an extremely difficult task, what shows the works such as [8, 9]. Therefore, it is not sufficient to use such a small number of parameters (3-4), for the systems that will be exposed to an adverse conditions and the noise level is considerably high.

There are also studies of slightly different manner. For instance, it is worth mentioning the work [10], which combines the power of statistics with the ideas of edge detection in computer vision. Another approach [11] uses no linguistic knowledge, but rather machine-learning algorithms based on clustering and dynamic programming techniques. In [12], a noise adaptive speech recognition system is built with acoustic models trained on noisy data, which is not common to the traditional approaches, where the training set is a clean speech.

The analysis of all works shows that cepstral-based approaches rightfully became popular among speech recognition community, but the need for more robust representation suggests augmenting them with additional cues such as acoustic parameters. Although another approaches like [6, 12] are possible to overcome noisy environment, we focus, more, on proper description of a speech signal. Here we do not discuss any work related to HMM; one may refer to [13, 14].

III. DATA ANALYSIS

Among the commonly used acoustic parameters we selected and analyzed four computationally easy to extract parameters: a maximum energy location (frequency), an energy concentration up to 1 kHz, Wiener entropy and a zero crossing rate of the autocorrelation function. Figure 1 shows the distribution of the parameters across the phonemes within each sonority class.

The location of the maximum energy (ME) in the spectrogram can be a cue in distinguishing between the sonorant and obstruent phonemes. Sonorant phonemes mostly have peak energy below 1500 Hz, however, the

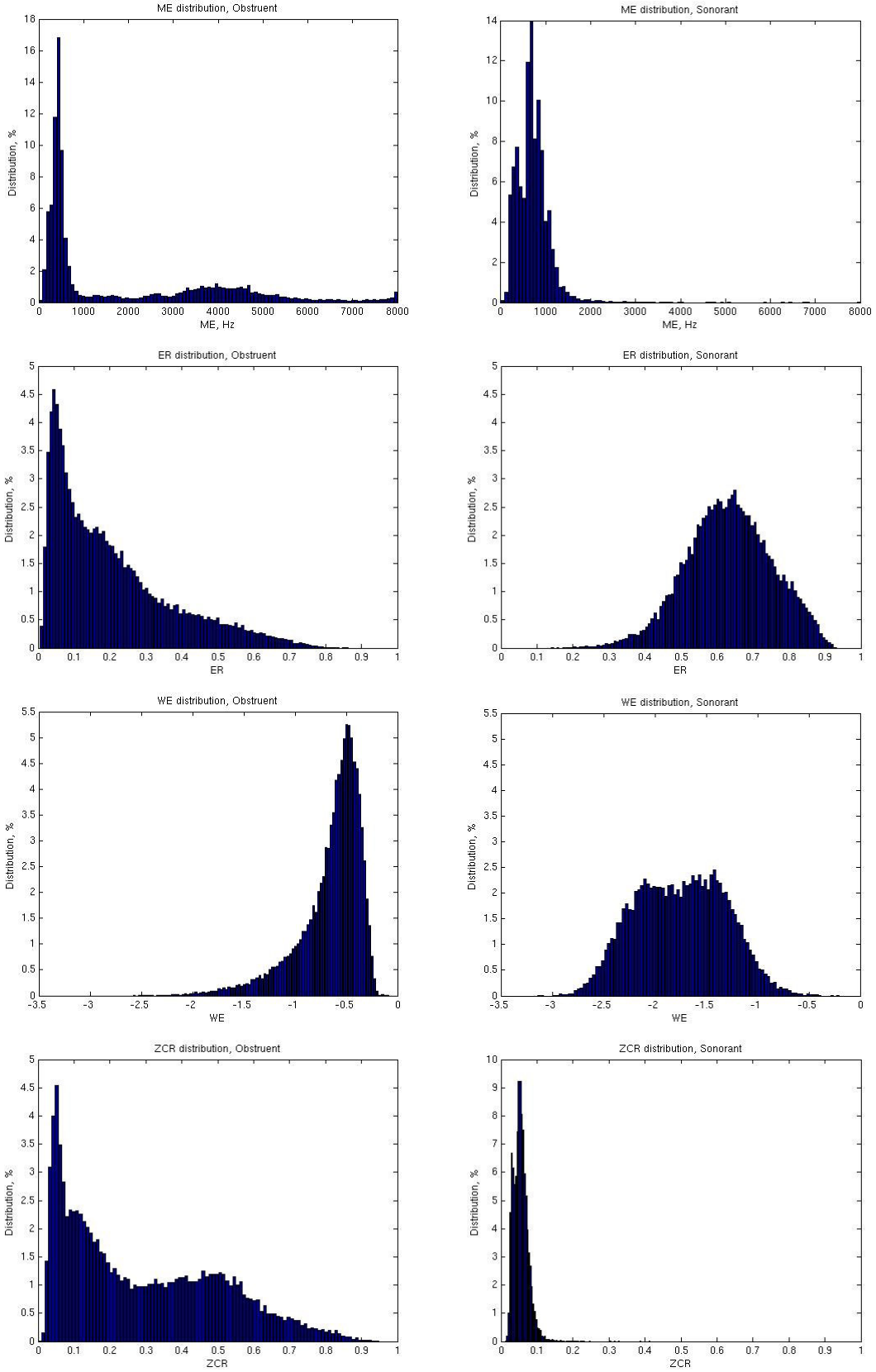


Figure 1. The distributions of the acoustic parameters across the obstruent (left) and sonorant (right) phonemes in the clean TIMIT samples.

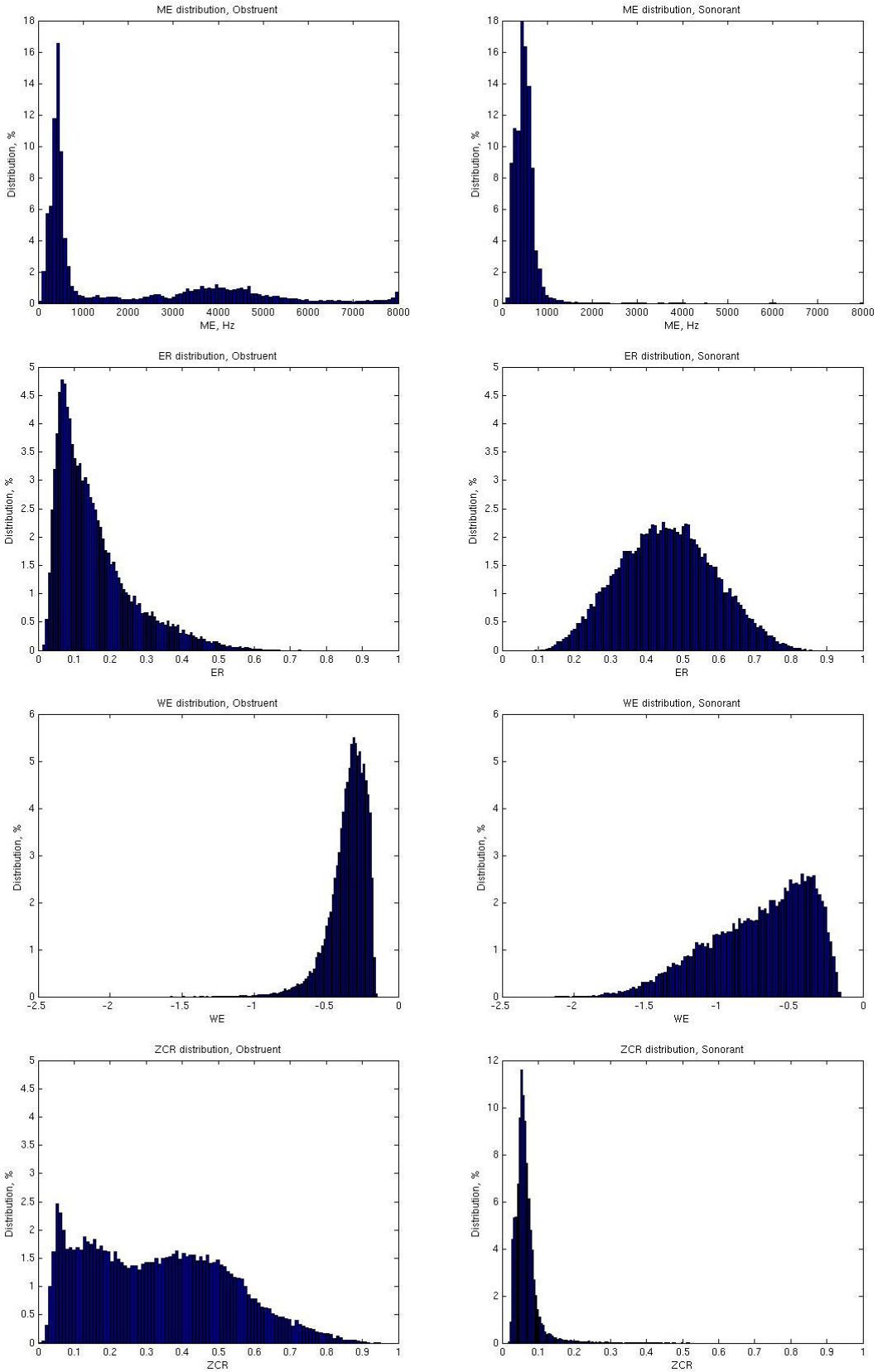


Figure 2. The distributions of the acoustic parameters across the obstruent (left) and sonorant (right) phonemes with normally distributed noise added (μ=0, σ=50, ~30dB).

range for obstruent phonemes varies much (see first row in Fig. 1). The vast majority of obstruent phonemes having low peaks are voiced phonemes like /b/ or /z/. Clearly, this parameter is good only in saying what is *not* a sonorant phoneme if the peak is high enough.

The second parameter, which we refer to as "energy ratio" (ER), is simply the ratio between the energy up to 1000 Hz to the total energy. The motivation for that is, again, the property of sonorants to have most of the energy in the low frequencies, and the reverse is for the obstruent phonemes. Although this statistics nicely separates two classes (see the second row in Fig 1.), the intersection of two histograms is still significant.

Wiener entropy (WE) is a measure of the width and uniformity of the power spectrum and defined as:

$$\int \log(S(f,t))df - \log(\int S(f,t)df), \qquad (2)$$

where S(f, t) is the energy in time-frequency domain. Since the sonorants have their energies high and concentrated in the lower frequencies, the value of this parameter tends to be small and far from zero, whereas the obstruent phonemes have their energies uniformly spread along the frequency axis up to 8000 Hz, what creates the flatness in power spectrum and, hence, pulls the entropy value towards zero. This can be seen in the Fig 1.

Similar measure, in the attempt of reflecting the structure, is a zero crossing rate (ZCR), the number of times the signal crosses zero to the total number of samples. It is an indirect measure, which reflects more-or-less the periodic structure of a signal. The sonorants being periodic will have this value low and the obstruents as well as noise will have it high. In addition, it should be noted that the value of a period for the sonorants is relatively high, due to the low frequency range, what is also a reason why ZCR is low for the sonorants. In our experiments we use the ZCR of the autocorrelation function, because the autocorrelation function has a smooth shape and theoretically the same period as the initial signal. Again, as in the case with the maximum energy, sonorants grouped compactly, while the obstruents do not show nice order, moreover, they fall also into lower "sonorant" region. The analysis shows that this intersection is caused by voiced obstruent phonemes, which have quasi-periodic shape.

In addition, we constructed the same distributions as in Fig.1 but with the normally distributed noise added, having mean $\mu=0$ and standard deviation $\sigma=50$, which roughly corresponds to ~30dB (see Fig. 2). Although the discussion of these diagrams will be made in the next section, it is worth noticing the significant changes in some of them, for instance, ER and WE. On the other hand, Figures 1 and 3 tell us that only two of four parameters, namely, ER and WE, sound promising to be used for separating the sonority classes. So, the important question here is how robust these parameters are if noise is added to the data.

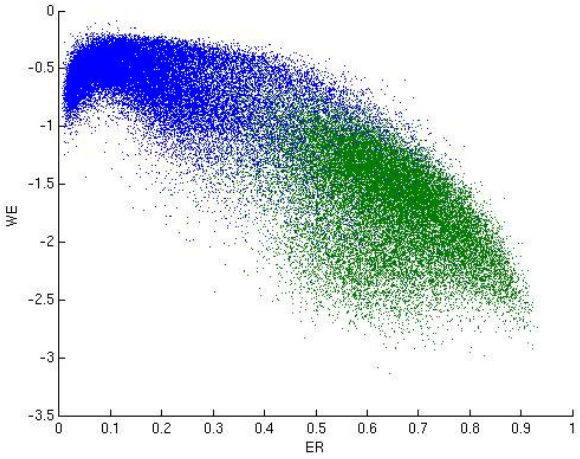
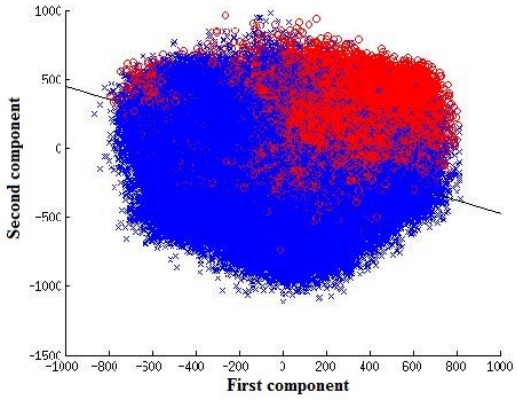


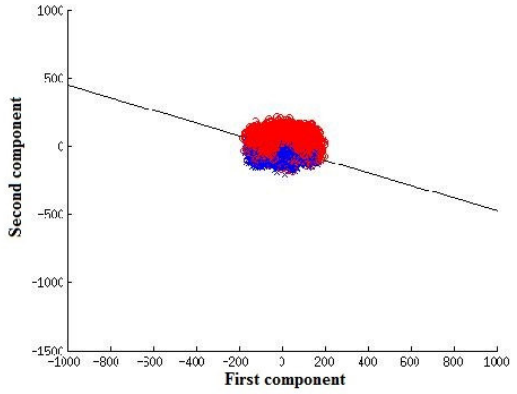
Figure 3. The distribution of the phonemes in the ER-WE plane (green points are sonorants, blue - obstruents).

The same analysis was made with MFCC features. For that, the MFCC vectors extracted from all the training samples were mapped using Principal Components Analysis (PCA) to a new space, where they were separated by the hyper-plane obtained as a result of Fisher's linear discriminate analysis (FLDA) [17]. Figure 4 shows the distributions of the vectors extracted from clean samples (a) and from noisy samples with noise level of 0dB (b) in the plane of the first two principal components (round markers correspond to the sonorant vectors, crosses – to the obstruent). Noisy data were transformed to the coordinate system of the clean data via PCA transform matrix. The lines on the graphs are the separating lines computed by FLDA. The change in the distribution is noticeable. Although it is not shown here, but the same shrinkage is seen along the other axis. Figure 5 compares the dynamics of the distribution of the vectors along the first principal component in the different noise environments.

Table 1 shows the results of the FLDA test for the other statistics. Here, we reflect how the separability of data changes by each parameter at noise level of 0dB. The first row is the separation error rate (% of wrong separated data out of all data) of the clean data obtained by FLDA, the second row is the separation error rate (with respect to the FLDA on clean data) of the noisy data, and the last row is the separation error rate obtained by retrained FLDA performed on the same noisy data. From this table we see that in general the separability of data decreases, though, it is still high for MFCC, ME and ER. Another interesting thing is that the error rate is high if we try to predict from the settings obtained on clean data, whereas if we retrain the system on noisy data the error rate decreases showing the better separation.



(a)



(b)

Figure 4. The distributions of the MFCC vectors in the plane of the first two principal components: a) the vectors were extracted from clean samples; b) the vectors were extracted from noisy samples with noise level of ~0dB (round markers correspond to the sonorant vectors, crosses – to the obstruent).

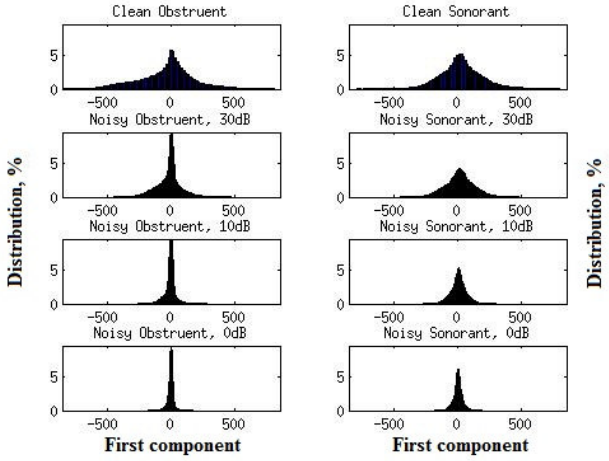


Figure 5. The distributions of the MFCC vectors along the first principal component in the different noise environments.

Table 1. The separation error rate, in %, obtained by FLDA for each parameter in clean and noisy environment.

Error rate, %	ME	ER	WE	ZCR	MFCC
clean	28	10	39	28	13
noise, 0db	21	52	52	60	20
retrained on noisy data	26	25	49	44	18

IV. EXPERIMENTAL RESULTS

4.1. Baseline system

To have a proper baseline with a reasonable performance, we selected the work done in [2] and tried to reproduce its results on the same 100 samples. We have built in Matlab 7.5 (R2007b) and HMM Toolbox [15] the HMM-based model using GMM as the observation probability density function with analogous settings. From each "sx" and "si" train utterances of TIMIT database, we extracted 13 mel-frequency cepstral coefficients every 10 ms over a 25 ms Hamming window, with subtracted cepstral mean computed over the all frames. In addition, we computed the deltas and the acceleration coefficients of the obtained MFCC vectors to capture their dynamic properties. Thus, we got a 39 dimensional cepstral feature set per utterance. Then each frame was labeled using the transcriptions according to the broad class it belonged to: vowels, semivowels, nasals, semivowels, fricatives, stops or silence. The extraction of the cepstral information was done using the MFCC Toolbox written by Daniel P. W. Ellis [16]. Expectation-Maximization (EM) algorithm was used to train the system.

As a primary metric to assess the performance of the model, we use one that is described in [2], namely:

1. Cson = percentage of the individual sonorant phonemes for which at least a fraction Fmin of its duration falls into a single sonorant segment as determined by algorithm.

2. Cobs = percentage of the individual obstruent phonemes for which at least a fraction Fmin of its duration falls into a single obstruent segment as determined by algorithm.

The rationale of this measure is that no matter how precise the phonetic transcription of a speech signal is, human is still able to identify the given phoneme even if we shrink a little the exact boundaries of it. Of course, the measure of "shrink a little" has its own limit depending on the characteristics of the phoneme and the perceptual ability of a human. Nonetheless, since we build our approach on the phonetic transcription of the utterances, we need to be aware of this phenomenon and permit to vary the acceptance level for the phonemes, in our case the value of Fmin. However, the most relevant values of Fmin are around 0.5, which means that we guess at least the center of a phoneme. Another positive side of this measure is that it reflects the quality of the segmentation: a "good" segmentation will

have both Cson and C_{obs} balanced, i.e. the difference between them is low. We explain it with a simple example. Suppose the algorithm predicted the whole signal as one big sonorant region, then, of course, we would have C_{son} = 100%, whereas C_{obs} = 0%, for all values of F_{min}. Clearly, this is not what we desire to achieve. On the other hand, exact match of the prediction with the real transcription would make both C_{son} and C_{obs} equal to 100%, which is ideal. So, generally for a "good" segmentation the difference $\Delta=(C_{son}-C_{obs})$ should be small and close to constant for all values F_{min} in [0,1]. Therefore, we prefer this measure to the commonly used string edit distance, which doesn't have such nice properties.

Table 2 compares both systems for the different values of F_{min}.

Table 2. Performances of the system done in [2] and our baseline system

F_{min}	System built in [2]		Our baseline system	
	C_{son}, (%)	C_{obs}, (%)	C_{son}, (%)	C_{obs}, (%)
0.10	98.5	95.4	97.5	91.7
0.33	96.6	92.9	96.1	90.6
0.50	95.0	89.3	93.4	89.2
0.67	93.4	85.8	88.9	86.9
0.90	82.1	68.7	71.3	77.6

4.2. Performance in clean and light noise environments

We made 7 different combinations of feature sets and tested them with our system on all testing samples to see their behavior in clean and light noise conditions (~30dB). The testing is performed on the all 1344 test utterances of "sx" and "si" type in TIMIT database. There are two purposes of these experiments: 1) to see how the performance changes if we add various acoustic parameters to the general-purpose MFCC-based feature set; 2) to estimate the robustness of these combination in the presence of light noise. The combinations are:

- MFCC (39 dimensional feature set, baseline)
- ER + WE (2)
- MFCC + ER + WE (41)
- MFCC + ME + ER (41)
- MFCC + ME + ER + WE (42)
- MFCC + ME + ER + WE + ZCR (43)
- MFCC + ME + ZCR (41)

The results are given in Table 3 and 4. For each system, we show the obstruent and sonorant prediction rates, as described in previous section. The important thing to notice in these numbers is not only the values C_{son} and C_{obs} but also the corresponding difference $\Delta=C_{son}-C_{obs}$ for a particular value of F_{min} as well as its dynamics: the smaller the values and the more uniform the change, the better performance of a system is.

From Table 3 we can see that the performance for the third system (MFCC+ER+WE) much better than that of the first one with MFCCs only. However, Figure 6 shows that the second (ER+WE) and the last (MFCC+ME+ZCR) systems have more-or-less stable behavior in term of evolution of Δ. The system with all the parameters included also seems stable for most of the values of F_{min}.

Table 3. The performance of the system with different feature sets in clean environment.

Fmin	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
MFCC										
C_{obs}, (%)	90.30	89.85	89.28	88.69	87.91	86.88	85.49	82.87	76.37	61.86
C_{son}, (%)	97.83	97.30	96.39	95.28	93.59	91.14	87.68	82.32	72.21	54.69
ER + WE										
C_{obs}, (%)	91.50	90.37	89.12	88.02	86.54	84.58	81.85	77.91	69.78	55.41
C_{son}, (%)	94.86	94.38	93.85	93.19	92.45	91.22	89.02	84.52	74.35	57.40
MFCC + ER + WE										
C_{obs}, (%)	94.46	94.15	93.74	93.20	92.43	91.04	88.94	85.08	77.14	61.47
C_{son}, (%)	98.00	97.52	96.84	95.82	94.29	91.63	87.78	81.70	71.09	55.03
MFCC + E + ER										
C_{obs}, (%)	89.47	88.63	87.60	86.64	85.60	84.09	82.09	79.11	73.17	62.03
C_{son}, (%)	95.46	94.78	93.80	92.63	90.79	87.08	82.04	75.03	65.52	53.66
MFCC + ME + ER + WE										
C_{obs}, (%)	86.74	85.75	85.03	84.45	83.67	82.60	81.08	78.25	72.85	62.61
C_{son}, (%)	94.32	93.45	91.97	90.26	87.91	84.62	80.13	73.52	63.81	52.62
MFCC + ME + ER + WE + ZCR										
C_{obs}, (%)	92.53	92.10	91.36	90.31	88.79	86.68	83.73	79.46	72.27	60.53
C_{son}, (%)	95.46	94.81	93.78	92.45	90.68	88.42	85.31	80.04	70.17	55.89
MFCC + ME + ZCR										
C_{obs}, (%)	92.15	91.53	90.54	89.23	87.60	85.22	82.00	77.58	65.93	53.48
C_{son}, (%)	95.59	95.03	94.09	92.80	91.14	88.99	86.01	81.18	72.31	58.36

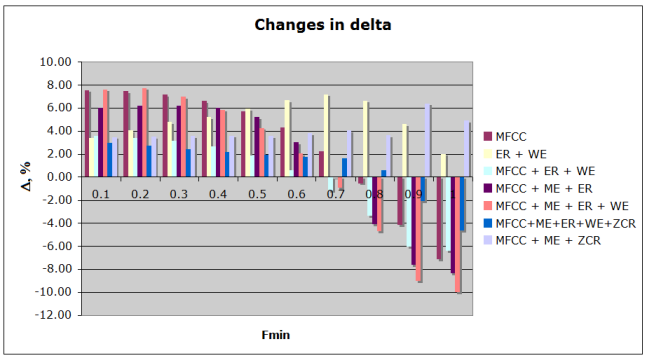


Figure 6. The dynamics of $\Delta = C_{son}-C_{obs}$ for each system in clean environment.

If in the clean conditions we can see comparable performance of all the systems to that of the baseline (MFCC), then in noisy conditions only few can compete with the baseline system (Tab. 4). There is a significant shift

in performance from C_{son} to C_{obs}, due to the fact that noise has the properties of the obstruent phonemes. On the other hand, it is interesting that the last system not only outperforms the baseline but also shows some stability of Δ value. The forth has lower performance than the last's but still somewhat stable. Figure 7 shows the change in Δ value only for the best 3 systems, which we chose to be the candidates for the next set of experiments.

Table 4. The performance of the system with different feature sets in noisy environment.

Fmin	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
MFCC										
C_{obs}, (%)	93.43	93.13	92.68	92.20	91.51	90.56	89.23	87.38	82.58	69.24
C_{son}, (%)	92.51	91.19	89.64	87.86	85.53	82.77	78.84	73.12	62.70	45.04
ER + WE										
C_{obs}, (%)	99.82	99.81	99.80	99.79	99.73	99.62	99.53	99.35	98.97	95.70
C_{son}, (%)	23.29	22.59	21.82	20.97	19.88	18.44	16.60	13.69	9.62	3.81
MFCC + ER + WE										
C_{obs}, (%)	97.56	97.49	97.34	97.19	96.94	96.46	95.61	94.21	91.36	82.32
C_{son}, (%)	84.36	82.50	80.14	77.38	74.21	70.33	64.67	56.75	44.57	29.00
MFCC + ME + ER										
C_{obs}, (%)	90.97	90.28	89.27	87.82	86.38	84.55	82.34	79.17	73.59	62.98
C_{son}, (%)	93.52	92.61	91.43	89.90	87.71	84.00	78.79	71.99	62.59	50.41
MFCC + ME + ER + WE										
C_{obs}, (%)	93.99	93.50	92.83	92.30	91.84	91.30	90.68	89.55	86.98	79.07
C_{son}, (%)	82.29	80.53	78.18	75.21	71.39	67.16	62.05	54.90	44.96	33.91
MFCC + ME + ER + WE + ZCR										
C_{obs}, (%)	96.44	96.30	96.10	95.79	95.39	94.59	93.38	91.61	88.00	78.96
C_{son}, (%)	81.80	79.91	77.53	74.76	71.45	68.04	64.06	58.32	48.32	33.93
MFCC + ME + ZCR										
C_{obs}, (%)	92.44	91.93	91.29	90.13	88.78	86.55	83.48	79.00	66.99	53.50
C_{son}, (%)	94.59	93.64	92.22	90.49	88.68	86.43	83.53	79.21	70.76	56.38

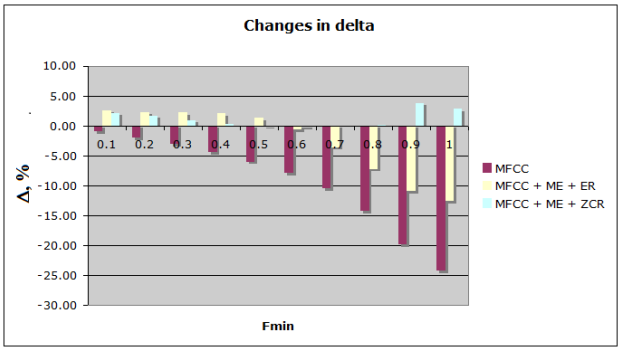


Figure 7. The dynamics of $\Delta = C_{son} - C_{obs}$ for the best 3 systems in noisy environment.

4.3. Performance in heavy noise

Based on the results of the previous experiments, we chose three best feature sets, which we tested in different noisy conditions. As a noise we used normally distributed noise with different means and standard deviations corresponding to various signal-to-noise ratios (SNR); the parameters are shown in the Table 5.

Table 5. The parameters of normally distributed noise.

SNR, dB	Mean, μ	Standard deviation, σ
25	50	50
20	0	100
10	0	300
0	0	800

The results are given in the Tables 6-9. A careful reader may notice that although there is more and more shift towards obstruent measure as more noise increased, the shifts for the systems with APs are relatively lower than that of the baseline. And even when all three systems fail in the noise of 0dB, the sonorant rate is still comparably high for the last system – twice as much as baseline's.

Table 6. The performance of the best candidates in noise of 25 dB

Fmin	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
MFCC										
C_{obs}, (%)	93.16	92.88	92.49	92.08	91.48	90.56	89.28	87.33	82.42	69.19
C_{son}, (%)	92.56	91.32	89.73	87.91	85.77	82.85	78.80	73.23	63.16	45.20
MFCC + ME + ER										
C_{obs}, (%)	92.93	92.51	92.03	91.22	90.18	88.48	86.23	83.11	77.85	67.80
C_{son}, (%)	90.75	89.68	88.23	86.36	83.81	79.84	74.29	67.10	57.53	45.19
MFCC + ME + ZCR										
C_{obs}, (%)	92.41	91.94	91.31	90.37	89.24	87.48	84.87	81.26	70.65	58.15
C_{son}, (%)	92.35	91.23	89.53	87.64	85.43	83.10	79.86	75.06	66.35	51.49

Table 7. The performance of the best candidates in noise of 20 dB

Fmin	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
MFCC										
C_{obs}, (%)	95.06	94.89	94.56	94.23	93.70	92.96	92.01	90.63	87.13	76.02
C_{son}, (%)	83.88	82.14	80.26	78.13	75.48	72.36	68.23	61.83	51.45	34.62
MFCC + ME + ER										
C_{obs}, (%)	92.15	91.57	90.64	89.55	88.34	86.44	84.17	80.95	75.33	64.71
C_{son}, (%)	88.82	87.58	85.77	83.62	80.98	77.31	71.94	65.05	56.23	44.59
MFCC + ME + ZCR										
C_{obs}, (%)	92.93	92.61	92.04	91.27	90.18	88.25	85.48	81.31	69.51	55.95
C_{son}, (%)	91.50	90.30	88.54	86.32	84.13	81.71	78.74	74.35	65.74	51.47

Table 8. The performance of the best candidates in noise of 10 dB

Fmin	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
MFCC										
C_{obs}, (%)	98.03	97.94	97.79	97.65	97.45	97.12	96.68	96.16	94.82	88.82
C_{son}, (%)	49.95	48.29	46.38	44.33	41.95	39.12	35.36	30.33	23.15	13.13
MFCC + ME + ER										
C_{obs}, (%)	96.57	96.32	95.99	95.68	95.14	94.32	93.02	90.97	87.17	78.44
C_{son}, (%)	55.57	53.89	51.65	49.19	46.52	43.29	39.52	34.91	28.92	20.80
MFCC + ME + ZCR										
C_{obs}, (%)	96.30	96.05	95.77	95.43	94.94	94.00	92.56	89.96	80.63	69.07
C_{son}, (%)	67.53	65.87	63.79	61.49	59.15	56.82	54.01	50.29	42.94	29.84

Table 9. The performance of the best candidates in noise of 0 dB

Fmin	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
MFCC										
C_{obs}, (%)	99.66	99.64	99.59	99.53	99.47	99.35	99.18	98.95	98.65	95.63
C_{son}, (%)	14.04	13.29	12.57	11.74	10.88	9.71	8.34	6.75	4.59	1.98
MFCC + ME + ER										
C_{obs}, (%)	99.36	99.32	99.23	99.15	99.06	98.93	98.62	98.26	97.48	93.79
C_{son}, (%)	13.33	12.50	11.68	10.69	9.79	8.86	8.04	6.96	5.34	2.89
MFCC + ME + ZCR										
C_{obs}, (%)	99.22	99.18	99.10	98.98	98.82	98.51	98.09	97.53	91.83	86.41
C_{son}, (%)	23.62	22.54	21.36	20.20	19.22	18.06	16.90	15.03	11.54	6.02

V. DISCUSSION

We start our discussion from the histograms in the Figures 1 and 2. They answer to some questions about the acoustic parameters as well as pose new questions concerning them. First of all, from Figure 1 we can see that some of the statistics (like ER or WE) have the property of separating two sonority classes and the others don't (ME or ZCR). Also from both figures we notice another property that if the noise is added to a speech signal some statistics shift along the axis and change their shapes a lot while some don't do this much. As a result, these two properties directly or indirectly affect the performance and the robustness of the system.

The effect of separating property can be, clearly, seen for the second and third systems used in clean environment. The performance of the second system is comparable with that of the baseline yet only two features were used, and the performance of the third is the best among all systems. The answer to our initial hypothesis – whether or not the combination of the parameters is good – would be obvious if there were no noise. The same systems immediately fail once they are exposed to noise. However, the other statistics that don't show this nice separating property but somehow immune to noise, i.e. don't change much in noisy condition, happen to increase the robustness of a system if used together with something that has this separating property. The examples are the fourth and the last systems, which, indeed, exploit the power of both feature sets.

The experiments show that it is this "shifting" property what makes the systems to have a tendency to classify most of the phonemes to the obstruent class as more noise is added. In fact, it is this statistics that "shift" not the phonemes itself, i.e. the representation of data is poor. Many of the research works based on the landmark detectors rely on the fact that the there is a chance to at least capture the essence of a phoneme such as nuclei of the syllable or the closure-burst transitions of the stops [2, 5, 6] and they work fairly well. Since human can do this tasks more-or-less robustly, there must be a "good" representation of a speech signal, for instance, the peak of maximum energy (ME) proved to be such a representation. Figure 8 shows the histograms of the distributions of APs for both classes in the noise level of 10dB. Notice the drastic changes in most of the statistics.

As for the MFCCs, they are nothing but the same kind of statistics as APs and have their own "limit of use". The shrinkage shown in Fig. 4 and 5 is an evidence of that. The last set of experiments shows slow degradation of the performance accompanied with the gradual shift of the performance towards the obstruent measure, and at the noise level of 0dB it shows that these features become useless. Although the AP statistics can be more-or-less robust to noise, the whole system fails because of this "limit of use" of the MFCCs. So, the question remains open if there is an alternative to MFCCs, which has longer "limit of use" and has nice separating property.

Another important question is that how reasonable the model is. It turns out that the well-known GMMs, which try to locate the objects in some space, in fact, are powerless, when those object start "shifting". Not only do they shift but also change their shape, i.e. the distribution in that space. The same problems will experience all models that are based on the "static" properties of data unless a "better" representation of it is found or they take into account its "dynamic" properties. Table 1 clearly shows that there can be some gains if the model can adaptively change its settings with respect to noise level.

Summarizing, we restate that there is a trade-off between separating quality and robustness of the acoustic parameters, and depending on the environment used respective feature should be granted. As far as noise concerned, since it affects the behavior of the features causing changes in shape and position in a given space, the statistical models should take into account these dynamics in order to be applicable in the adverse conditions

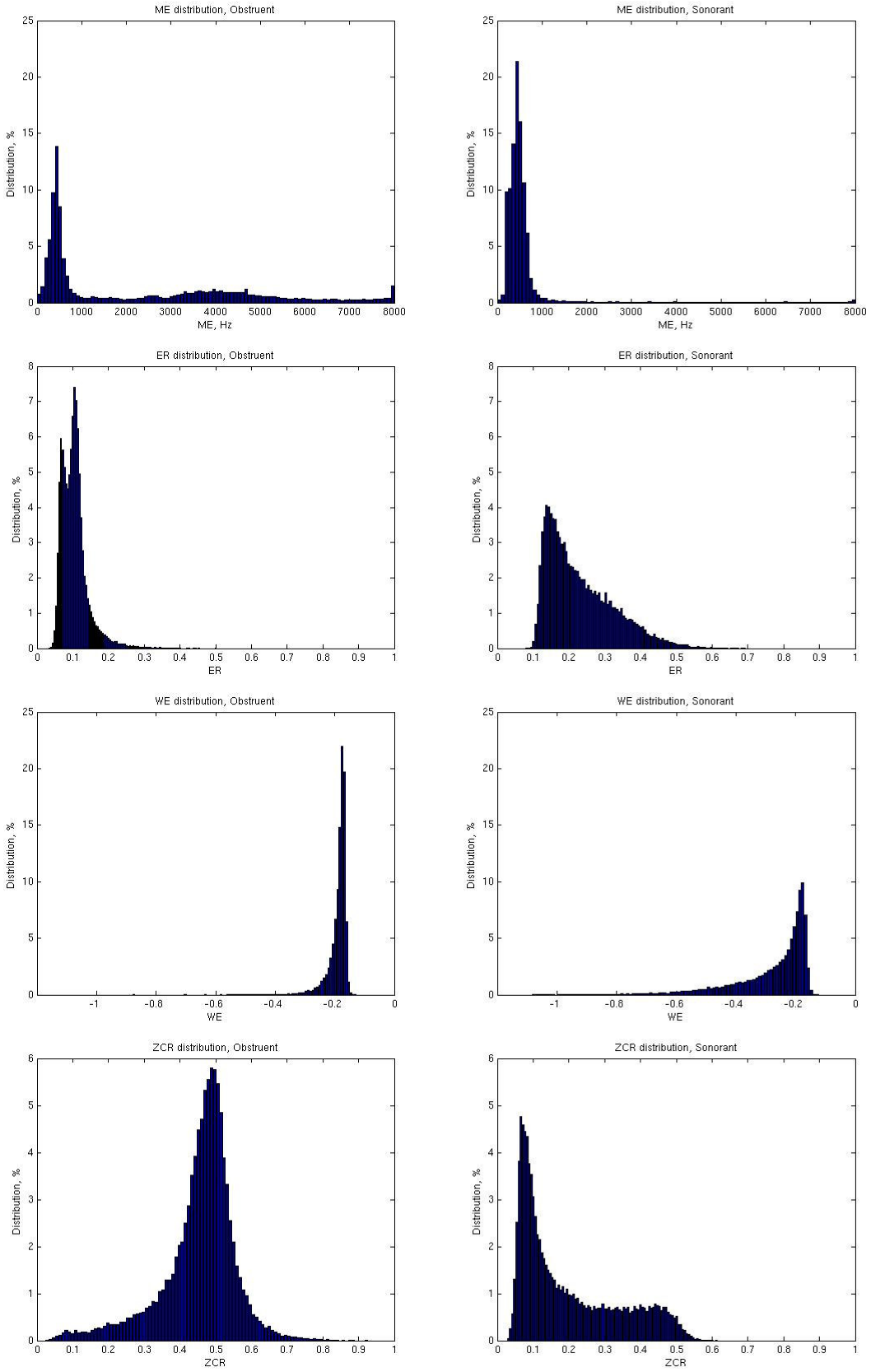


Figure 8. The distributions of the acoustic parameters across the obstruent (left) and sonorant (right) phonemes with normally distributed noise added (μ=0, σ=300, ~10dB).

VI. CONCLUSION

In this paper, we analyzed several acoustic parameters to see how robust they are in the noisy conditions and estimated their performance combined with cepstral coefficients in the task of segmentation of continuous speech into sonorant and obstruent regions. The results show that if the "dynamics" of the analyzed statistics in noise is taken into account, one can achieve better performance of a system.

As a future work it is planned to investigate some other acoustic parameters and build an adaptive system that would adjust to changing nature of the parameters in noise.

REFERENCES

[1] J. P. Olive, A. Greenwood, J. Coleman, *Acoustics of American English speech. A dynamic approach*, Springer, 1993

[2] A. Jansen, P. Niyogi, "A probabilistic speech recognition framework based on the temporal dynamics of distinctive feature landmark detectors," *Tech. Report TR-2007-07*, 2007

[3] K. Schutte, J. Glass, "Robust detection of sonorant landmarks," *Proceedings of Interspeech*, pp.1005-1008, 2005

[4] A. Juneja, C. Espy-Wilson, "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," *Proceedings of 9th International Conference on Neural Information Processing*, Singapore, Volume 2, pp. 726-730, 2002

[5] P. Niyogi, C. Burges, and P. Ramesh, "Distinctive feature detection using support vector machines," *Proceedings of ICASSP,* pp.425-428, 1999

[6] Zhimin Xie, P. Niyogi, "Robust acoustic-based syllable detection," *Proceedings of Interspeech*, paper 1327-Wed1BuP.6., 2006

[7] P. Niyogi, E. Petajan, J. Zhong, "Feature based representation for audio-video speech recognition," *Proceedings of the Audio Visual Speech Conference*, pp. 133-139, 1999

[8] S. Parthasathy, S. Mehta, S, Srinivasan, "Robust periodicity detection algorithms," *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 874–875, 2006

[9] M. Vlachos, Philip Yu, V. Castelli, "On periodicity detection and structural periodic similarity," *In Proc. of SIAM International Conf. on Data Mining (SDM)*, pp. 449-460, 2005

[10] Y. Amit, A. Koloydenko, and P. Niyogi. "Robust acoustic object detection," *J. Acoust. Soc. Am*, 118(4), pp. 2634-2648, 2005

[11] M. Sharma, R. Mammone, "Blind" speech segmentation: automatic segmentation of speech without linguistic knowledge," *In ICSLP-1996*, pp. 1237-1240, 1996

[12] Kaisheng Yao, K.K. Paliwal, S. Nakomura, "Noise adaptive speech recognition with acoustic model trained from noisy speech evaluated on Aurora-2 database," *In ICSLP-2002*, pp. 2437-2440, 2002

[9] Lawrence R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, 77 (2), pp. 257–286, 1989

[10] Jeff A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models," *Tech. Report TR-97-021*, Berkeley, CA, April 1998

[15] Kevin Murphy, HMM Toolbox, 1998
http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

[16] Daniel P. W. Ellis, PLP and RASTA (and MFCC, and inversion) in Matlab, 2005 [Accessed: April 20, 2011].
http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/
[Accessed: May 15, 2011].

[17] Fisher, R., "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7, pp.179–188, 1936