Evaluation of Simultaneous Speech Detection Based on MFCC-DTW with Two-Stage Normalization

Alexandru-George Rusu[†], Radu-Sebastian Marinescu[*], Corneliu Burileanu[*], and Dumitru Bica[†]

***Abstract*— In Air Traffic Control a serious safety risk is represented by undetected simultaneous transmissions from different airplanes. In this paper, we approach this issue through a speech analysis algorithm, which combines traditional Mel Frequency Cepstral Coefficients extraction, a new two-stage normalization and widely used Dynamic Time Warping. In this way, we were able to extend the simultaneous speech detection capability in Voice Communication Systems of Air Traffic Control. The results prove that this implementation is suitable for practical applications.**

***Keywords*—Dynamic Time Warping; Mel Frequency Cepstral Coefficients; speech analysis; two-stage normalization.**

I. Introduction

In the last decades technology evolved in an explosive way. This also happened with the communications and transportation capabilities. As a consequence, the air traffic was also constantly increasing and more and more planes operate now in the same time in the air and on the ground. Thus, in such high density operational airplanes areas, the chances that two, or even more planes, to initiate a communication with the air traffic controller is also increasing. When this happens and the controller is not aware of both transmissions we have a high safety risk. These situations are acknowledged by aeronautical authorities and threated as a potentially safety issue to be resolved.

Therefore, solutions as in [1] were proposed to alert the air traffic controller when one ground radio station receives simultaneous radio signals from at least two airplanes (e.g. the scenario shown in Fig. 1). This approach uses radio spectrum analysis for a single receiving channel of several emitters. However, this solution, as presented by aeronautical standard [2], resolves only a part of the problem. The case when two simultaneous emissions from two airplanes are received separately, by two different ground radio stations, cannot be threaded as in [1] using radio spectrum analysis for a single channel and from our knowledge, no solution was communicated. This is because the traffic controller will hear only one received signal by a channel (usually selected by Received Signal Strength Indication – RSSI), the system discarding the unselected signals. Each received radio signal will contain mainly only the emission from the nearest airplane and at the control office will be no information that the received signals came from different source. Thus, we developed a new approach which uses together all receiving signals in a multichannel speech information analysis. In this way, the system will be aware if simultaneous different transmissions are received by different radio stations.

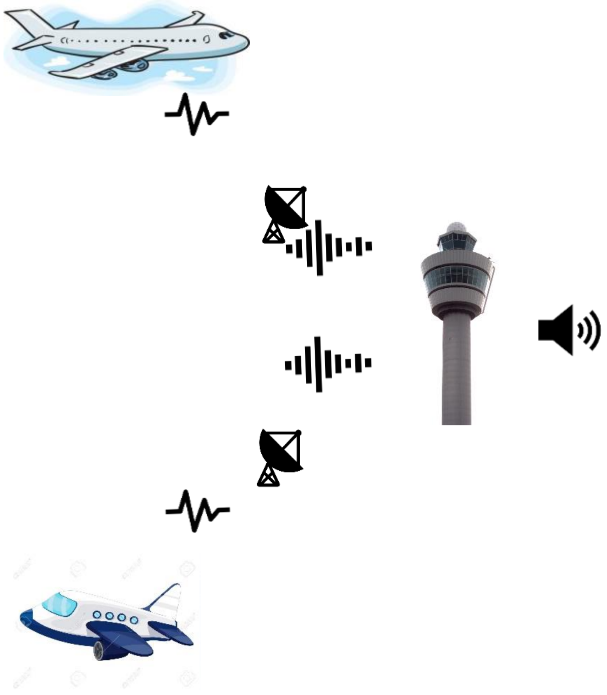

Fig. 1. Scenario of simultaneous transmissions on the same frequency

A remark must be pointed to the specific aspects and scenarios of these voice communication systems for air traffic control which make this task more difficult. The length of analyzed signals is maximum 300ms, due to real time communications demanding. Beside this, we have to deal with inconstant delays on each reception. Another thing which hinders our task is the noise level variation, due to automatic gain control. This last factor can alter efficiency when stabilization of automatic gain control (AGC) takes more than 200ms. In Fig. 2 we have the representation of the same emitted signal, received by three radio station. The

Manuscript received October 31, 2018, revised May 14, 2019.

[†]Alexandru-George Rusu and Dumitru Bica are with Research and Development Department, Rohde&Schwarz Topex, Bucharest, Romania (email: {alexandru.rusu, dumitru.bica}@rohde-schwarz.com)

[*] Radu-Sebastian Marinescu and Corneliu Burileanu are with Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Bucharest, Romania (email: radu.marinescu@speed.pub.ro, corneliu.burileanu@upb.ro)

doi: 10.11601/ijates.v8i2.278

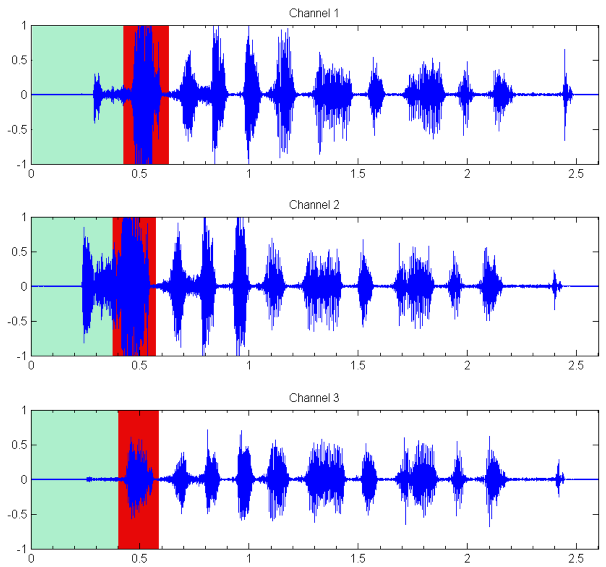

Fig. 2. Received signals; the red zone, which was to be used for simultaneous call detection, contains the first uterance, which is affected by different noise levels on each channel

green area represents the first part of the received signals, where there is noise and the speech is difficult to be heard.

In a straightforward way, in our case we have to verify if two, or more received signals came from the same pilot, transmitting the same voice message, or came from different pilots with different meanings. After a pre-processing stage, with a specifically tuned voice activity detection derived from [3] and possible used of a time delay estimation block [4], our problem looks like a speaker recognition task.

In this approach, usually in the first stage are extracted speech features, like widely used, Linear Predictive Coding (LPC) [5], Mel Frequency Cepstral Coefficients (MFCC) [6] or Perceptual Linear Predictive coefficients (PLP) [7]. Then, the evaluation makes use of Dynamic Time Warping (DTW) [8], or Hidden Markov Models (HMM) [9], depending on each system's specifics.

Next, the paper is organized as follows: Section II contains details of our previous and actual approach. In Section III we present experiments regarding tuning and configuration of our method, while evaluation and discussions are reserved for Section IV. Conclusion and further work stand in section V.

II. Simultaneous Speech Detection

A. Speech Feature Based Attempts

Due the fact that our working signals have a limited duration of less than 300ms, are affected by delay and varying noise level, we first tried to extract some robust speech features, used also in forensics domain. We evaluate in different configuration and combinations the Mean Delta [10], Entropy-Energy [11], log windowed autocorrelation lag energy (logWALE), spectral autocorrelation peak to valley ratio (SAPVR), modified-SAPVR [12]. We also extracted pitch information, using well known algorithms, such as PRAAT [13], RAPT [14], SRH [15], and YIN [16].

Then, the DTW was applied for various combinations of the above extracted speech features. The obtained results were not accurate enough to be used in practice. An explication for this fact could be found in the high noise level presented on the first part of the received signals. However, it is worth mentioning that in some cases, when the AGC gets stabilized quick, we can achieve reliable results.

B. 2-Stage normalization MFCC and DTW solution

In order to improve previous results we also apply DTW for MFCC. This approach has been also used before in various tasks [17-20]. On our first implementation in this way, we obtain much better results, compared with those from previous attempts, but not enough accurate.

For further enhancement, what we propose is a two stage normalization. Then, the first normalization is done by computing the z-score (or standard score) to MFCC from each separate 1 frame, as follows:

$$MFCC_{l,i}^{n1} = \frac{MFCC_{l,i} - \mu(MFCC_l)}{\sigma(MFCC_l)}, \qquad (1)$$

where $\mathrm{MFCC}^{n1}_{l,i}$ represents the *i*-th first normalized MFCC of frame *l*, while μ and σ stand for mean and standard deviation. This could be seen as a normalization along frequency. The variance and mean at this stage, are computed along the MFCC dimension. The second normalization takes place in the cost function of the DTW, computing the standardized Euclidean distance between the two MFCC-vectors from each frames of each signal, with the variance now computed along the time series. This could be seen as normalization along time, because it operates on values from different frames.

III. Experiments Preparation

For experiments, we evaluate several scenarios using different data base, TIMIT speech corpus [21], Noisex-92 noise corpus and a proprietary database with recorded signals from voice communication systems for air traffic control. Because in our systems the sampling frequency is set to 8kHz, we downsampled the reference signals from TIMIT and Noisex-92. For simulations, we used Python programming with its specific modules (e.g. Scipy, Numpy, Pandas).

To get insight of how the noise and delay affects our system, we separated our experiments in two stages. First, we configured and evaluated our algorithm on the clean TIMIT database, inserting artificial delays. After choosing an optimum configuration, we proceeded to the noisy analysis. In this last experiment, we selected several types of noises from Noisex-92 and added to TIMIT signals to obtain different SNR levels from -5dB to 15dB. In this way we could perform a second calibration phase. Finally, we check our proposed algorithm with the real database from operational field, which is affected by heavily variant noise and delays.

When analyzing the effects produces by delay, we are interested in delays from 0 to approximately 100ms. This range is used as a safety measure, because in voice communications systems for air traffic control, based on VoIP, maximum accepted delays by network's QoS are usually less than 40ms. We selected 2342 clean signals from TIMIT database and grouped in pairs of two, one with each other, adding artificially generated delay to the second signal from each pair. Using this combination, we obtain 2342*2342 pairs, from which 2342 pairs should be detected as the same transmission (which are formed from the same signal), and the rest as different speech.

For MFCC and DTW computation we set frame size to 128 samples with a step of 64. Also, in order to obtain fast results as imposed by [2], we used a fixed signal length of 300ms, which corresponds to 37 frames. The other parameters had optimally been chosen to increase accuracy, as show in Table I. Then, we extracted the discrimination dependency over the MFCC number.

We decreased the number of missed and false detection using two normalization stages (first in frequency and second in time). The results of simulations on clear pairs signals, with different normalizations at relevant delays are presented in Table II. For each delay and normalization configuration combined 2342 signals. This yields 2342x2342 pairs, of which 2342 should be detected as *same speech*, and the others as *different* or *simultaneous speech*. Missed detections count undetected different speech pairs, which were labeled erroneously, as same speech. False detections count erroneously detected pairs of same speech as different speech.

After a first calibration step, we were interested in how the system's accuracy is influenced by the number of MFCC. Tring to minimize the number of missed and false estimations we obtain an optimum 33 MFCC, in range 8 to 40, which was used in the next simulations.

TABLE I
PARAMETER SELECTION

Parameter	**Min Value**	**Max Value**	**Optimum**
Pre-emphasys coeff	0	1	0.97
No_filter	26	40	max(30, no_MFCC)
Window	none	Hamm, Hann	none
Celplifter	0	22	0
Delta MFCC	0	20	0
Delta-Delta MFCC	0	20	0
Append_energy	False	True	True
First normalization	by frequency	by time	by frequency
Second normalization	by frequency	by time	by time

TABLE II
MISSED AND FALSE DETECTIONS

Delay [ms]		**72**	**80**	**96**
Missed detections	No Normalization	23	2794	374 176
	Freq. normaliz.	54	4062	591 349
	Time normaliz.	0	0	37
	Both normaliz.	**0**	**0**	**1**
False detections	No Normalization	7	60	524
	Freq. normaliz.	8	49	557
	Time normaliz	0	0	19
	Both normaliz	**0**	**0**	**0**

IV. EXPERIMENTS EVALUATION AND DISCUSSIONS

A. *Delay and noise influence on Timit database*

In the calibration phase, using only clean signals, for more than 20 MFCCs we obtained 100% accuracy for delays bellow 50ms. This can be observed in Fig. 3 for the false rate. Equivalent results were obtained also for missed rate.

When adding several types of noise (white, pink, HF, engine or cockpit), without introducing any delay, the missed and false rate increased. However, these rates can be adjusted by a detection threshold. Rising detection threshold will decrease the false detection rate, but will increase much more the missed detection rate. In air traffic control, these rates should be small as possible. Because of this, a balance must be chosen, to have as less as possible false detections, while maintaining a reasonable detection rate for simultaneous different speech. This is because the impact of a false detection to the traffic controller is higher than in the case of a missed detection.

Fig. 4 shows how the false rate is affected by the combination of delays and white noise at -5dB SNR. The threshold has been varied between 1.25 and 1.40 and the lowest false rate is achieved for a threshold of 1.40.

For comparison, in Fig. 5 is presented the effect of delays and noise at 10dB. In both figures we can also notice that using a higher detection threshold reduces the false rate.

The characteristics of missed rate increases also with delay, as shown in Fig. 6, for noisy signals. In this scenario, the threshold of 1.40 leads to the highest missed rate.

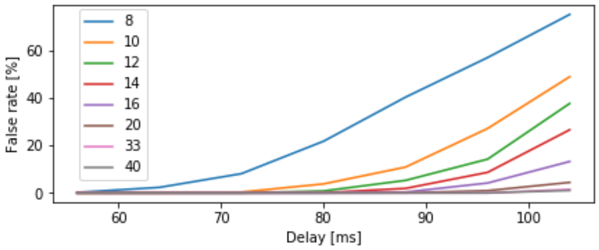

Fig. 4. Dependency of false rate by numbers of MFCC, clean signals

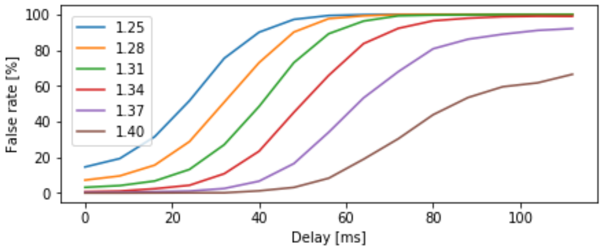

Fig. 5. False rate [%] with white noise, SNR -5dB, 33 MFCCs

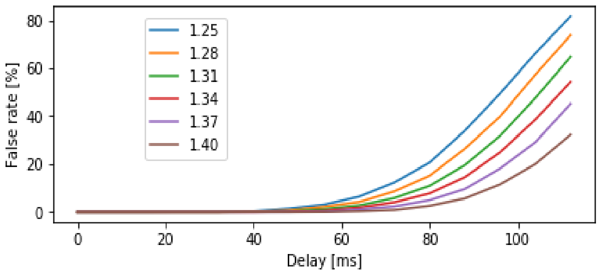

Fig. 6. False rate [%] whit white noise, SNR 10dB, 33 MFCCs

However, the missed rate could be drastically reduced in practical scenarios for relative low delays by tuning the threshold decision on lower values. For noisy signals and delays bellow 50ms, a good tradeoff between missed and false rate can be find using a threshold between 1.30 and 1.35.

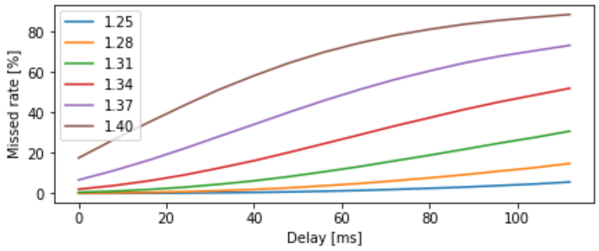

Fig. 7. Missed rate [%] whit white noise, SNR 5dB, 33 MFCCs

TABLE III
MISSED RATE [%] FOR DIFFERENT NOISES AND SNR

Noise type SNR [dB]	white	HF	F16	pink
-5	0.54	0.49	0.12	0.45
0	0.55	0.49	0.12	0.44
5	0.87	0.83	0.45	0.51
10	0.87	0.83	0.46	0.52
15	0.87	0.83	0.46	0.52

TABLE IV
FALSE RATE [%] FOR DIFFERENT NOISES AND SNR

Noise type SNR [dB]	white	HF	F16	pink
-5	2.22	1.49	4.9	2.55
0	1.62	1.07	5.2	2.49
5	0	0	0	0
10	0	0	0	0
15	0	0	0	0

TABLE V
MISSED RATE [%] FOR DIFFERENT NOISES AND SNR, 16MS DELAY

Noise type SNR [dB]	white	HF	F16	pink
-5	2.34	2.10	1.87	2.02
0	2.39	2.10	1.87	2.01
5	3.01	2.78	1.93	2.01
10	3.09	2.78	1.95	2.46
15	3.09	2.78	1.95	2.45

TABLE VI
FALSE RATE [%] FOR DIFFERENT NOISES AND SNR, 16MS DELAY.

Noise type SNR [dB]	white	HF	F16	pink
-5	4.43	1.40	6.10	4.56
0	2.53	1.20	7.23	4.59
5	0	0	0	0
10	0	0	0	0
15	0	0	0	0

Hence, in Table III and Table IV we show the results for different noise types at different SNR, with the threshold 1.32, chosen in the middle part of the tradeoff interval. We used this value to avoid *false detections* at positive SNR. For SNR at 0 dB and below the signals are heavily affected by noise, resulting in an increasing number of *false detections*. Regarding *missed rate*, in this configuration, it appears to be stable for positive SNR values. However, for negative SNR we obtained reduced values for it. This could be explained by an increase of the DTW score with the increase level of noise. Thus, several DTW values of different pairs, which were below the threshold, now are detected correctly as different speech. Nevertheless, the decreasing speed of missed rate is significantly lower than increasing speed of false rate.

In Table V and VI are summarized the results for different noise types at different SNR and a delay of 16ms, with the threshold 1.32, chosen in the middle part of the tradeoff interval.

B. Analysis on Operational Signals

In our previous simulations we used a combination of clean and noisy signals, the last ones with relatively stable and known SNR. These configurations let us tune algorithm for several aspects and gave insight about its robustness. Moving forward, after calibration stage, we proceed with new experiments using now operational signals. Compared with our artificially prepared noisy signals, the operational ones do not have a stable noise level. The noise level is usually higher at the beginning of the speech. After the AGC is stabilized, which typically takes 30 to 300ms, the SNR is high and we could say that we have proper signals to use on. However, because of safety measures and standard limitations, it is not allowed to wait more than 300ms. Thus, processing must start on heavily noise affected signals, sometimes being also distorted. Because of this, the results for operational signals are worse than from previous experiments.

For our current analysis we used 360 signals, from 3 radio channels, which were grouped in 120 pairs. With tuned configuration from earlier experiments we evaluated our algorithm on all operational signals. As we expected, missed and false rate have opposite characteristics and a trade-off must be chosen. In Fig. 7 presented their trends. It must be noticed that the dynamic range of false rate expands almost

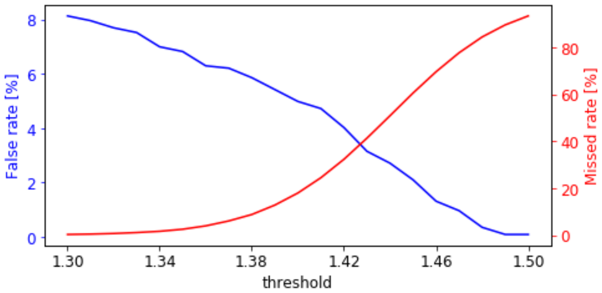

Fig. 3. Missed and false rate dependency by decision threshold, for operational signals

on its entire domain, while the missed rate remains below 10%, when using a decision threshold between 1.3 and 1.5. Another remark is that around this interval, an amount of false rate decrease implies 10 times increase to missed rate.

In order to mitigate both errors in the same time, we added several pass-band filters. In the Tables VII, VIII, IX are shown the results obtained in the scenarios where a band pass filter (BPF) has been used and compared with the result summarized in Table X without BPF.

In the tests involving the band pass filter, we varied both the lower and upper frequencies with a resolution of 50 Hz between 300Hz and 3300Hz. In the noisy scenarios it was remarked that the use of the band pass filter leads to lower missed and false rate, but when the SNR presents high values (e.g. 10dB, 15dB), the band pass filter leads to worse results.

On the signals from the TIMIT database this measure decreased the missed rate detection for heavily affected noise signals, but increased both rates for less noise affected signals. Also, different low pass filters have been used where the low pass frequency has been varied between 300Hz and

TABLE VII.
FALSE RATE [%] AND MISSED RATE [%] WHEN A BPF IS USED (LOW PASS FREQUENCY: 300HZ, HIGH PASS FREQUENCY: 1500HZ)

Delay [ms]	Missed Rate [%]	False Rate [%]
0	1.31	6.06
8	2.77	5.22
16	4.90	4.71
24	7.72	4.55
32	11.70	4.38
40	16.19	4.21
48	21.33	3.87
56	27.80	3.87

TABLE VIII.
FALSE RATE [%] AND MISSED RATE [%] WHEN A BPF IS USED (LOW PASS FREQUENCY: 300HZ, HIGH PASS FREQUENCY: 2000HZ)

Delay [ms]	Missed Rate [%]	False Rate [%]
0	1.031	5.556
8	2.226	5.387
16	3.736	4.882
24	6.061	5.051
32	8.967	5.051
40	12.229	4.714
48	16.636	4.882
56	21.89	4.54

TABLE IX.
FALSE RATE [%] AND MISSED RATE [%] WHEN A BPF IS USED (LOW PASS FREQUENCY: 300HZ, HIGH PASS FREQUENCY: 2000HZ)

Delay [ms]	Missed Rate [%]	False Rate [%]
0	0.72	5.89
8	1.44	5.56
16	2.56	5.39
24	4.12	5.39
32	6.20	5.22
40	8.65	4.71
48	12.12	4.71
56	16.00	5.05

TABLE X.
FALSE RATE [%] AND MISSED RATE [%] WITHOUT BPF

Delay [ms]	Missed Rate [%]	False Rate [%]
0	0.52	5.39
8	1.08	5.05
16	1.91	4.71
24	3.14	4.71
32	4.64	4.71
40	6.52	4.38
48	8.88	4.21
56	11.81	4.21
64	15.20	4.71

TABLE XI.
FALSE RATE [%] AND MISSED RATE [%], MFCC – ZCR

Delay [ms]	Missed Rate [%]	False Rate [%]
0	24.11	4.20
8	27.52	3.70
16	31.37	3.36
24	34.90	3.03
32	38.63	3.36
40	42.68	3.19
48	46.75	3.03
56	51.07	2.86
64	54.81	2.37

TABLE XII.
FALSE RATE [%] AND MISSED RATE [%], MFCC – ZRMSE

Delay [ms]	Missed Rate [%]	False Rate [%]
0	24.11	9.09
8	26.32	9.60
16	27.38	9.09
24	28.54	9.26
32	29.62	9.76
40	30.43	9.60
48	31.15	10.27
56	31.99	9.93
64	31.59	10.77

1200Hz, but the results were the similar as in scenarios with band pass filter.

Another approach consists in the mix between MFCCs and zero crossing rate (ZCR) or combination between the zero crossing rate (ZCR) and the short-term energy (called as ZRMSE) as the last coefficient. It was expected that these features would lead to high errors at low SNRs, but tables X and XI show that this fact is true for high SNRs.

The errors introduced in the case when ZCR or ZRMSE have been used have increased considerably.

V. Conclusion and Further Work

In this paper, we proposed an algorithm to detect simultaneous speech from voice communications systems for *air* or *naval* traffic control, which is based on traditional Mel Frequency Cepstral Coefficient and Dynamic Time Warping. We introduced a 2-stage normalization procedure, which reduced missed and false detection rates. We also analyzed the use of other several speech features, simulated with TIMIT, Noisex-92 and a proprietary data base.

On future experiments we will mixt the MFCC with various speech features aiming for better results. We will also try different new combinations adding speech features presented in Section II.A. A new research direction for this task is represented by deep learning, which recently was used in [22] for overlapped speech detection.

References

[1] L. Friedrich, "Method and device for the detection of simultaneous dual emission of AM signals", pattent DE102007037105 A1, 2008.

[2] ED-137B - "Interoperability Standards for VoIP ATM Components", European Organization for Civil Aviation Equipment, 2016.

[3] R. S. Marinescu, C. Burileanu, "Voice activity detection for best signal selection in air traffic management and control systems", in the *Proc. 38th International Conference on Telecommunications and Signal Processing (TSP)*, Prague, Czech Republic, 2015.

[4] R. S. Marinescu, "Best Signal Selection with Automatic Delay Compensation in VoIP Environment", PhD Thesis, University Politehnica of Bucharest, Romania, 2013.

[5] T. E. Tremain, "The government standard Linear Predictive Coding Algorithm: LPC10", *Speech Technology*, Vol. 1, No.2, pp. 40-49, 1982.

[6] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed., pp. 374–388. Academic, US, 1976.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of Acoustical Society of America*, no. 87, pp. 1738-1752, 1990.

[8] R. Bellman, "Dynamic Programming," *Princeton University Press*, 1957.

[9] M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.

[10] A. Ouzounov, "Robust Feature for Speech Detection", *Cybernetics and Information Technologies*, vol.4, No.2, pp.3-14, Bulgaria, 2004.

[11] L. S. Huang and C. H. Yang, "A novel approach to robust speech endpoint detection in car environments", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1751-1754, Turkey, 2000.

[12] T. Kristiansson, S. Deligne, P. Olsen, "Voicing Features for Robust Speech Detection", in *Proc. 6th Annual Conference of the International Speech Communication Association (ISCA)–INTERSPEECH*, pp. 369-372, Portugal, 2005.

[13] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings. Institute of Phonetic Sciences*, University of Amsterdam, pp. 97-110, 1993.

[14] D. Talkin, "Speech Coding and Synthesis", *Elsevier Science B.V.*, 1995.

[15] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, Italy, 2011.

[16] A. de Cheveign´e and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917-1930, 2002.

[17] V. Ramasubramanian, A. Das, and V. Kumar, "Text-dependent speaker-recognition using one-pass dynamic program-ming", in *Proc. ICASSP'06*, France, 2006.

[18] N. Murali Krishna, P.V. Lakshmi, Y. Srinivas, J. Sirisha Devi, "Emotion Recognition using Dynamic Time Warping Technique for Isolated Words", *International Journal of Computer Science Issues*, Vol. 8, Issue 5, No 1, pp. 306-309, 2011.

[19] S. V. Chapaneri, "Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping", *International Journal of Computer Applications*, vol, 40, no. 3, pp. 6-12, 2012.

[20] W. Fu, X. Yang, and Y. Wang, "Heart sound diagnosis based on DTW and MFCC", *3rd International Congress on Image and Signal Processing*, pp. 2920-2923, 2010.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscuc, D. S. Pallett, N. L. Dahlgren, V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1", Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[22] V. Andrei, H.Cucu, C. Burileanu, "Detecting overlapped speech on short timeframes using deep learning", *INTERSPEECH 2017*, pp. 1198-1202, Sweden, 2017.