Comparison of skewness-based salient event detector algorithms in speech

Annamária Kovács[1,2], Gábor Kiss[2], Klára Vicsi[2], István Winkler[1,3], Martin Coath[4]

[1]Research Centre for Natural Sciences
Hungarian Academy of Sciences
Budapest, Hungary

[2]Department of Telecommunication and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary

[3]Institute of Psychology
University of Szeged
Szeged, Hungary

[4]Cognition Institute
Plymouth University
Plymouth, United Kingdom

***Abstract*—In this work, we compare two skewness-based salient event detector algorithms, which can detect transients in human speech signals. Speech transients are characterized by rapid changes in signal energy. The purpose of this study was to compare the identification of transients by two different methods based on skewness calculation in order to develop a method to be used in studying the processing of speech transients in the human brain. The first method, the skewness in variable time (SKV) finds transients using a cochlear model. The skewness of the energy distribution for a variable time window is implemented on artificial neural networks. The second method, the automatic segmentation method for transient detection (RoT) is more speech segmentation-based and developed for detecting transient-speech segment ratio in spoken records. In the current study, the test corpus included Hungarian and English speech recorded from different speakers (2 male and 2 female for both languages) Results were compared by the F-measure, the Jaccard similarity index, and the Hamming distance. The results of the two algorithms were also tested against a hand-labeled corpus annotated by linguistic experts for an absolute assessment of the performance of the two methods. Transient detection was tested once for onset events alone and, separately, for onset and offset events together. The results show that in most cases, the RoT method works better on the expert labeled databases. Using F measure with +- 25ms window length the following results were obtained when all type of transient events were evaluated: 0,664 on English and 0,834 on Hungarian. Otherwise, the two methods identify the same stimulus features as the transients also coinciding with those hand-labeled by experts.**

Keywords—auditory events; speech transients; automatic speech segmentation; artificial neural networks; skewness; auditory feature extraction

I. INTRODUCTION

Transient intervals in speech are characterized by rapid changes in the overall energy of the signal or the distribution of energy in the frequency space. Transient extraction in speech is important for speech processing because it plays a fundamental role in speech sound identification and discrimination [1] [2]. In several speech studies, transients serve as a cue for finding the most informative sections of the speech signal [3] [4]. It is well documented that transients are enhanced by the auditory periphery and the mid-brain, and also a large part of the neurons in auditory cortex respond to rapid changes [5]. In human perception, transients are important for speech comprehension, object recognition, sound grouping, etc. Results showed in this work can serve as a sequel of [6] in the sense of early detection of health problems. According to the definition of Cognitive infocommunications (CogInfoCom) [7][8], transient detection can enhance the intra- and inter-cognitive communications. For these reasons, we believe that automatic transient detection is important in many aspects of the CogInfoCom area [7][8].

The development of the algorithm from a new aspect i.e. transient detection using artificial neural networks can bring closer transient detection task into a more biologically plausible implementation. In this work, we compare two methods of transient detection. The first (SKV) is a biologically plausible artificial neural network implementing a model of auditory transient extraction in the brain [9]. The second (RoT) was designed specifically to segment speech into transient and quasi-stationary parts based on special spectral distance. The latter method has been used for developing an early non-invasive biomarker for detecting possible health-related changes in the crew members of the Concordia Research Station [6].

978-1-4673-8129-1/15/$31.00 ©2015 IEEE

II. DATABASES

We used the speech record of four (2 male/2 female) native Hungarian speakers retelling in infant-directed speech a folk tale "The flying box", which was originally used for another experiment [10]. From the original recording, only 10 sentences were used for each speaker to have similar amount of signals as obtained from an English language corpus (40 sentences, altogether). The English corpus is a subcorpus of Stevens' Lexical Access From Features corpus [11]. Speakers of this corpus are native American or Canadian English speakers. Audio files from both corpora have been sampled at 16 kHz. Sentence length varies between 23 seconds. The signals for both languages have been labeled by experts for speech transients. The expert labels (EL) served as ground truth in assessing the efficacy of the methods.

III. METHODS

The performance of the two methods (SKV and RoT) was tested to identify transient intervals against EL data. We defined two types of transients: "onset" transient, where the overall energy is rising and "offset" transient where the overall energy is falling.

A. Skewness in variable time (SKV)

Speech sounds have been first processed with the Simple Cochlear model [13] using 128 Gammatone filters from 100 Hz to 8000 Hz center frequencies, which were arranged evenly on an equivalent rectangular bandwidth scale [14]. The skewness in variable time (referred as SKV) calculation [12] used the output of this model as the input to calculate the skewness of the distribution of the energy in a time window with a length of (1) and with a minimum time value of 3 milliseconds in high frequency bands. Thus the length of the window depends on the central frequency of the channel.

$$8 * 1/\mathrm{CF} \qquad (1)$$

The skewness is the third central moment of the distribution. It serves as a measure of asymmetry: is the amount by which a distribution is skewed to the right or to the left compared to the Gaussian curve. Right-shifted distributions (longer left 'tail') are categorized as 'onset' while left-shifted distributions (the right 'tail' is longer) are categorized as 'offset'. Skewness calculation was implemented on biologically plausible artificial neural networks [5], a two-layer feed-forward neural network with five sigmoid hidden neurons, used here as a function approximator. The traditionally calculated SKV values were used to train this network, where the inputs were the vectors of the values derived channel-by-channel from the Simple Cochlear Model. Levenberg-Marquardt backpropagation was used for implementing learning in the network. As in [12] the neural networks display compressive non-linearity showing relative insensitivity to the amplitude of the incoming signal, particularly at higher representation levels. This is a feature which is desirable in peripheral processing as it is a primitive form of representational invariance. After the SKV calculation for the 128 frequency bands the values across all channels were summed every 5 ms to produce a time series with 5 ms steps. After normalization, a threshold was applied to the series resulting in a binary event-series for the incoming speech signal based on the given threshold. Using the neural networks as function approximator was selected here both to achieve reduction in the computational cost and for biological plausibility.

B. Automatic segmentation method for transient detection (RoT algorithm)

This method was developed to measure the transient and quasi-stationary speech periods separately. [6].

The method is the following: First, frequency analysis of the incoming sound was performed using 30 milliseconds window length and 5 millisecond time step using a Hamming window. In the next step, the band-filtering of the signal was carried out using a Mel-band filter-bank from 300 Hz to 5200 Hz, resulting 20 bands. After this the segmentation of the speech signal is based on a spectral distance of two neighbouring intervals. Each interval was represented by the following variables: the first three Mel-bands with the largest mean intensity $I_{MB}{}^{x}{}_{k}$ (k=1,2,3), where x indicates the index of the *k*-th largest mean intensity of the Mel-bands and $I_{MB}{}^{x}{}_{k}$ is the intensity of the *x*-th Mel-band of the interval; the mean intensity of the interval (I_{mean}); and the variance of the mean intensity of the Mel-bands in the interval (I_V). The three distance measures of two neighbouring intervals *j* and *j-1* were calculated by the following formulae.

$D^{j}_{I_{mean}} = \sqrt{\left| I_{mean_j}{}^{2} - I_{mean_{j-1}}{}^{2} \right|}$, where I_{mean_j} is the I_{mean} of the interval j.

$D^{j}_{I_V} = \sqrt{\left| I_V{}_j^{2} - I_V{}_{j-1}^{2} \right|}$, where I_{V_j} is the I_V of the interval j.

$$D^{j}_{I_{MB}} = \frac{\sum_{k=1}^{3} \sqrt{\left| I_{MB_j}{}^{x_{j_k}}{}^{2} - I_{MB_{j-1}}{}^{x_{j_k}}{}^{2} \right|} + \sum_{k=1}^{3} \sqrt{\left| I_{MB_{j-1}}{}^{x_{j-1_k}}{}^{2} - I_{MB_j}{}^{x_{j-1_k}}{}^{2} \right|}}{6} \qquad (1)$$

, where x_{j_k} indicate the index of the *k*-th largest mean intensity of the mel bands of the interval j, and $I_{MB_j}{}^{x_{j_k}}$ is the $I_{MB}{}^{x}{}_{k}$ of the interval j. The distance of two neighboring intervals is defined by the weighted average of these distance measurements.

$D_j = \frac{w_1 D^{j}_{I_{mean}} + w_2 D^{j}_{I_V} + w_3 D^{j}_{I_{MB}}}{w_1 + w_2 + w_3}$, if the overall energy change was positive

$D_j = -\frac{w_1 D^{j}_{I_{mean}} + w_2 D^{j}_{I_V} + w_3 D^{j}_{I_{MB}}}{w_1 + w_2 + w_3}$, if the overall energy change was negative

At the beginning of the algorithm, the speech was segmented into 5 ms long intervals without clustering. The label of the first interval (transient or quasi-stationary) was determined on the basis of D_1. If D_1 was larger than the threshold, then it was classified as "onset" transient, if D_1 was smaller than minus one times of the threshold, then it was classified as "offset" transient otherwise it was classified as quasi-stationary. The threshold value (T_h) represented the big rapid changes which indicate transient parts. After the first interval was clustered, the algorithm worked iteratively. Interval *j* was

classified as "onset" transient if D_f (i.e., the distance of the actual undecided interval and the previous already clustered interval) was larger than T_h or "offset" transient if D_f was smaller than $-T_h$, otherwise quasi-stationary. If the actual segment got the same label as the previous segment, then they were merged. The algorithm was stopped after the last undecided 5 ms long interval was classified, thus organizing the segmentation of the speech into transient and quasi-stationary intervals [6].

The weights (w_1, w_2, w_3) and threshold value (T_h) was determined by getting the best F score value on the EL results discussed later on the section of the "Results".

C. Comparison methods

We first optimized the threshold values by comparing the results of each algorithm against the EL ground-truth data. Thresholds were applied to transform the series of SKV/distance values into binary transient/non-transient vectors. The timing of the transient was defined as the latency of the maximal SKV/distance value within the transient sections. First, only the onsets, then both the onsets and offsets were used in comparison between the two methods.

In binary classification tasks, such as the comparison of the SKV and RoT algorithms, precision and recall are the commonly used metrics to measure the quality of the results. Precision measures the relevancy of the results, while recall the returned truly relevant results. The combination of them, F-measure determines the accuracy of the comparison results in a value between 0 and 1. Small time-shifts are possible in the latency of transient maxima, therefore to eliminate the differences resulting from this, windowed F-measure calculation was applied with a maximal value of +-50 ms windows. Other comparison methods, such as Jaccard similarity index and Hamming distances were used as additional measures. Jaccard similarity index [15] [16] provides a value between 0 and 1 which indicates the similarity and diversity of the binary results sets. Hamming distance [17] is a well-known method to define the distance between two binary vectors, the less the value is, the more similar the datasets are. The windowed F-measure method with the additional Jaccard similarity index and Hamming distance are good indicators of correspondence between the results of the two algorithms.

IV. RESULTS

A. Comparison of the SKV and RoT algorithms with the EL database

First we compared the results of each algorithm (pooling both languages) against the EL results and determined the threshold values for the two methods. Threshold-values were optimized by the F-measure.

Tests were performed to get the best threshold value pairs for the SKV and RoT methods and the EL transient events, the best values were 30% for the SKV and 60 for the RoT method.

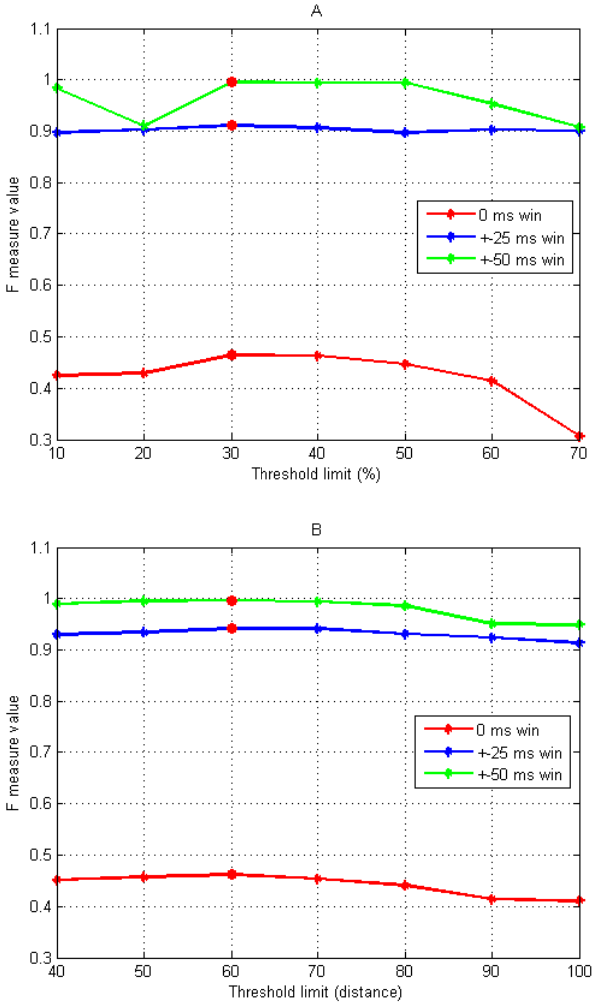

Fig. 1. F-measure values at three different thresholds (10, 20, and 30%) and window sizes (0, +-25, +-50 ms), separately for the SKV (Fig.1/A) and the RoT (Fig.1/B) method. Thin red dots are indicating the highest F-measure values, which have been used at further tests.

Fig. 1 shows the F-measure values computed to get the best threshold values, separately for the SKV and the RoT methods. It can be seen, that the SKV algorithm performs best at 30% threshold level: The F-measure is higher at 30% threshold than at the other surrounding thresholds with all three tolerance windows. For the RoT, the 60 threshold level proved to be the best.

B. *Comparison of the SKV and RoT algorithms using the best threshold values*

TABLE I. Onset transient events: F-measure values for three different window lengths at the best thresholds (SKV: 30%; RoT: 60) for the Hungarian and the English corpus using the EL as ground truth and the relative improvement (RI) of the RoT compared to the SKV

	0ms	+-25ms	+-50ms
English (SKV/EL)	0.120	0.562	0.809
English (RoT/EL)	0.248	0.664	0.902
The Relative improvement of the RoT compared to SKV	**106%**	**18%**	**11%**
Hungarian (SKV/EL)	0.160	0.578	0.928
Hungarian (RoT/EL)	0.355	0.834	0.965
The Relative improvement of the RoT compared to SKV	**121%**	**44%**	**4%**

TABLE II. Onset and offset transient events: F-measure values for three different window lengths at the best thresholds (SKV: 30%; RoT: 60) for the Hungarian and the English corpus using the EL as ground truth truth and the relative improvement (RI) of the RoT compared to the SKV

	0ms	+-25ms	+-50ms
English (SKV/EL)	0.240	0.833	0.962
English (RoT/EL)	0.313	0.874	0.978
The Relative improvement of the RoT compared to SKV	**30%**	**5%**	**2%**
Hungarian (SKV/EL)	0.465	0.911	0.995
Hungarian (RoT/EL)	0.462	0.942	0.996
The Relative improvement of the RoT compared to SKV	**-1%**	**3%**	**0%**

As indicated in Table I. and Table II. in all cases the RoT algorithm performed better, but it was developed to find transient-like sections in speech. When both onset and offset transients were included, there were no big differences between the F-measures for the two languages, but in the Hungarian language RoT performed slightly better in the case of onsets until the +-25 ms time window, at +-50 ms window length the performance is similar. This reason, we may claim, these algorithms perform almost the same for the both languages, and can detect automatically transients.

Using the best thresholds, we compared the two algorithms to find out whether both are finding the same transient events. Onset and onset/offset values are compared against each other in both languages as shown on Fig. 1. and 2. Additionally the Jaccard indices and Hamming distances were calculated (Table III., IV.) The highest F-measure values were achieved using +-50 millisecond window length, as at this window length, the SKV method incorporates more events, which are associated with transients detected by the RoT.

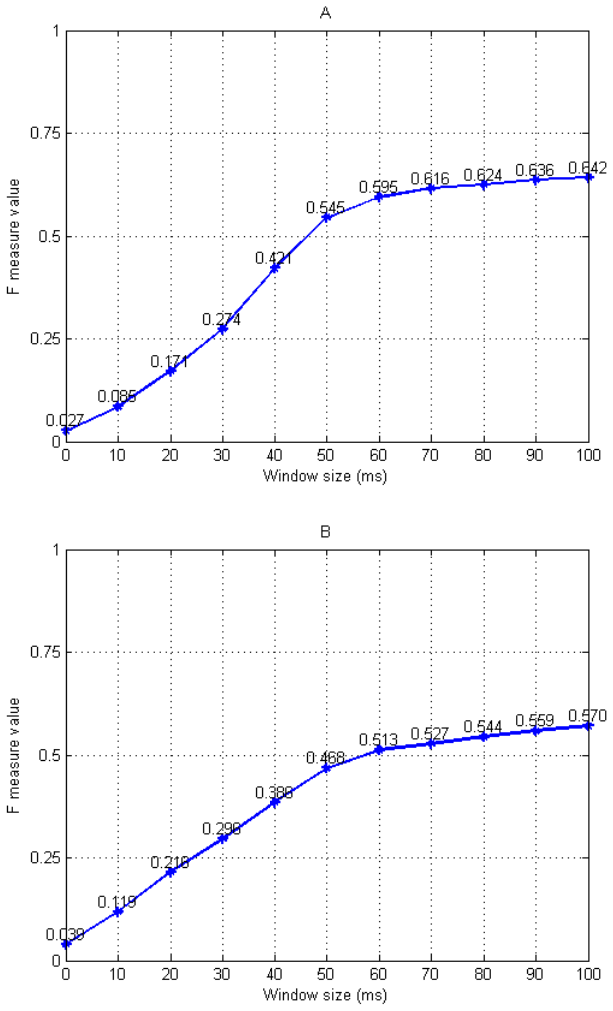

Fig. 2. F-measure values for comparing the two methods on the English corpus with different window lengths. Fig. 2/A shows the values based on the onset transients only, Fig. 2/B is based on both the onset and offset transients. There is a small difference between the two, because the offset transients are not as well approximated as the onsets.

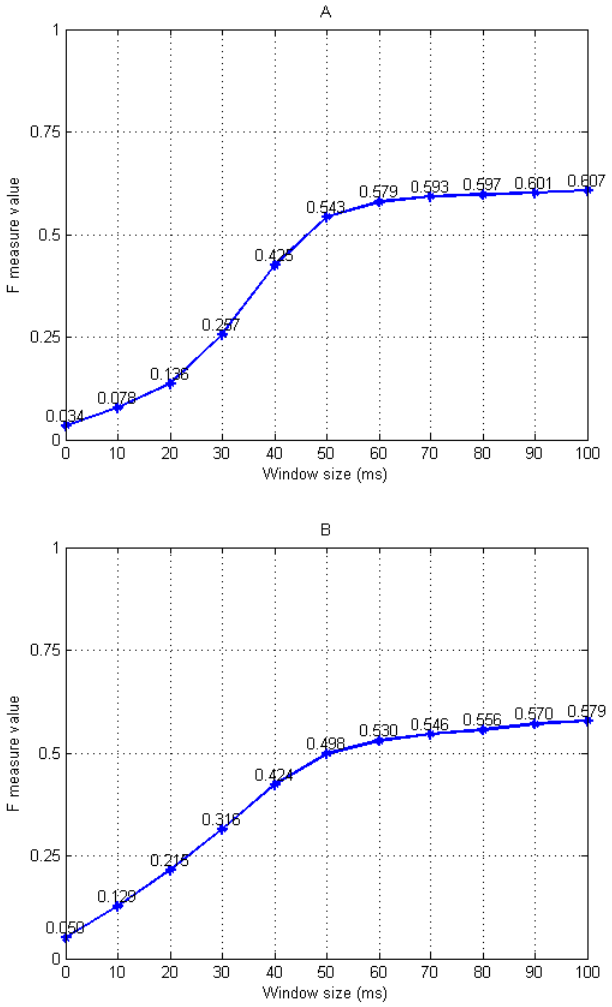

Fig. 3. F-measure values for comparing the two methods on the Hungarian corpus with different window lengths. Fig. 3/A shows the values based on the onset transients only, Fig. 3/B is based on both the onset and offset transients . There is a small difference between the two, because the offset transients are not as well approximated as the onsets.

As shown on the graphs, F-measure values are almost the same for the the Hungarian and the English corpus. It is growing sharply until 50 millisecond window value, after that it shows small only a small increase, and reaches its maximum at +-50 milliseconds. This means that temporal precision of the algorithms is approximately ±25ms.

TABLE III. Jaccard indices and Hamming distances calculated between the results of the two algorithms with different window lengths at the best thresholds (SKV: 30%; RoT: 60) on the English corpus.

ENG	Measure	0ms	20ms	50ms	70ms	100ms
Onsets	Jaccard	0.013	0.161	0.518	0.587	0.604
	Hamming	661	303	174	149	143
Onsets and offsets	Jaccard	0.020	0.198	0.440	0.503	0.538
	Hamming	1031	471	329	292	271

TABLE IV. Jaccard indices and Hamming distances calculated between the results of the two algorithms with different window lengths at the best thresholds (SKV: 30%; RoT: 60) on the Hungarian corpus.

HUN	Measure	0ms	20ms	50ms	70ms	100ms
Onsets	Jaccard	0.017	0.168	0.686	0.743	0.757
	Hamming	964	336	127	104	98
Onsets and offsets	Jaccard	0.026	0.247	0.590	0.650	0.695
	Hamming	1552	536	292	249	217

Table III. and IV. indicate that Jaccard index and Hamming distance together with the F-measure can serve as a good measure to compare the algorithm results. For the English corpus, the onsets match better than the combination of onsets and offsets. For the Hungarian corpus the same results were found using the F-measure, whereas the Jaccard indices are higher and the Hamming distances are lower, indicating the opposite.

V. Conclusion

Both algorithms produced phasic peaks that define transients. These transients emerge within short time windows, where there are changes in the spectral content. There are no large differences in the performance of the two algorithms for English and Hungarian speech, and both algorithms' results matching with the expert-labeled transients rather well, with slightly better values for the RoT algorithm. According to the results it can be stated that both algorithm can detect transients within 25ms precision time window, because until the +-25ms

window length the F measure values show significant increasing tendencies, nevertheless after that it is not remarkable. The RoT algorithm was designed to measure the clearness of the articulation of the speech, in contrast the SKV algorithm was designed for auditory event detection, although they worked almost the same efficiently on these databases. These results could lead us to implement a transient-predictor with the combination of the two algorithms, which can be used to find possible neural responses.

Finding transients in speech is an important task in the light of that these segments are carrying the most information in the speech signal. If the transient detection can be implemented automatically, it can help to understand these sections better, which can be really important in speech comprehension. The combination of the two above mentioned algorithms can help to solve this problem and detect the transients in a more biologically plausible way. Furthermore the better understanding and detecting transient parts of the speech may enhance the communication between different cognitive beings as for example it may help to differ between separate speech events and emotional states [18].

ACKNOWLEDGMENT

This work is funded by: Lendület Projekt (Magyar Tudományos Akadémia) 5-0105-2012-5112

REFERENCES

[1] B. Wildermoth, "Use of voicing and pitch information for speaker identification," M.S. thesis, School of Microelectronic Engineering, Griffith Univeristy, Australia, 2001. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] Rasetshwane, D.M.; Boston, J.R.; Li, C.-C.; Durrant, J.D.; Genna, G., "Enhancement of speech intelligibility using transients extracted by wavelet packets," *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on* , vol., no., pp.173,176, 18-21 Oct.

[3] Mathiak K., Hertrich I., Lutzenberger W., Ackermann H., "Encoding of temporal speech features (formant transients) during binaural and dichotic stimulus application: A whole-head magnetencephalography study", Cognitive Brain Research, Volume 10, Issues 1–2, September 2000, Pages 125-131.

[4] Coath M., Denham S.L., "The role of transients in auditory processing". Biosystems, 2007, 89(1-3):182–189.

[5] Delgutte, B., "Auditory neural processing of speech". The Handbook of Phonetic Sciences, Oxford: Blackwell, 1997, 507-538.

[6] Kiss G., Sztaho D., Vicsi K., Golemis A., "Connection between body condition and speech parameters - especially in the case of hypoxia, " *Cognitive Infocommunications (CogInfoCom), 2014 5th IEEE Conference on* , vol., no., pp.333,336, 5-7 Nov. 2014.

[7] Baranyi P., Csapo A., , Varlaki P., "An Overview of Resarch Trends in CogInfoCom". *Intelligent Engineering Systems (INES), 2014 18th International Conference on , vol., no., pp.181-186, 3-5 July 2014 doi: 10.1109/INES.2014.6909365.*

[8] Baranyi, P. , Persa, G. , Csapo, A, "Definition of Cognitive Infocommunications and an Architectural Implementation of Cognitive Infocommunications Systems". *World Academy of Science, Engineering and Technology, International Science Index 58 (2011), 5(10), 376 - 380.*

[9] Kovács A., Coath M., Denham S.L., Winkler I., Mády K., "Sparse processing methods for locating information-bearing points in continuous speech signal", unpublished.

[10] Kocsis Zs., Winkler I., Szalárdy O., Bendixen A., "Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: An event-related potential (ERP) study", Biol. Psychology, Volume 100, 2014 July, 20-33M.

[11] Stevens, K. N., "Toward a model for lexical access based on acoustic landmarks and distinctive features". J Acoust Soc Am, 2002, 111(4):1872–1891.

[12] Coath M., Denham S.L., "Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience". BiolCybern, 2005, 93(1):22–30.

[13] Slaney M., "Auditory toolbox documentation". technical report 45. Technical report, Apple Computers Inc. 1994.

[14] Glasberg B.R., Moore B.C., "Derivation of auditory filter shapes from notched noise data", Hear Res, 1990. 47(1):103–138.

[15] Jaccard P., "Lois de distribution florale". Bulletin de la Socíeté Vaudoise des Sciences Naturelles 38:67-130, 1902.

[16] Jaccard P., "The distribution of the flora in the alpine zone." New Phytologist 11(2):37-50, 1912.

[17] Hamming, Richard W., "Error detecting and error correcting codes", Bell System Technical Journal 29 (2): 147–160, 1950.

[18] Truong, Khiet P; Van Leeuwen, David A, "Automatic discrimination between laughter and speech".Speech Communication, 49,2,144-158,2007,Elsevier