U.P.B. Sci. Bull., Series C, Vol. 77, Iss. 4, 2015 ISSN 2286-3540

PERCEPTION STUDY INSPIRED METHOD FOR AUTOMATICALLY DETECTING THE NUMBER OF COMPETING SPEAKERS

Valentin ANDREI[1], Horia CUCU[2], Andi BUZO[3], Corneliu BURILEANU[4]

We present a statistical method for determining the number of competing speakers in an audio recording, inspired from a social experiment which is also described. We mixed a set of high quality audio samples with continuous speech produced by single speakers and we played the tracks to each human listener within our target group. The volunteers were asked how many simultaneous speakers were they able to count and how they obtained the response. We observed that while human subjects showed a correct detection ratio of 31%, the artificial method was able to achieve 66% accuracy. In addition we highlight the observations related to speaker counting methods made by the volunteers and we also analyze if the proposed algorithm can serve as a high performance voice activity detector in a LV-CSR operating in real time.

Keywords: selective auditory attention, voice activity detection, blind source separation, multi-speaker recordings

1. Introduction

Large vocabulary continuous speech recognition systems (LV-CSR) have become “good enough” to be used in activities like controlling a device or dictating a text message, where recognition errors do not have an important impact. However the accuracy of these systems is seriously affected by non ideal usage conditions: voices with accents, emotions, jargon, crowded or noisy places, etc. In this paper we focus on dealing with crowded places where multiple simultaneous speech sources are present, trying to get one step closer to understanding how the human brain is able to switch attention from one speaker to another.

One of the main “features” that we have for following the desired speaker in a conversation is binaural hearing. This enables us to localize the spatial origin of the sound and focus our attention towards it. Binaural hearing is widely described in technical research studies being referred as to binaural processing,

[1] University POLITEHNICA of Bucharest, Romania, e-mail: valentin.m.andrei@gmail.com
[2] University POLITEHNICA of Bucharest, Romania, e-mail: horia.cucu@upb.ro
[3] University POLITEHNICA of Bucharest, Romania, e-mail: andi.buzo@upb.ro
[4] University POLITEHNICA of Bucharest, Romania, e-mail: corneliu.burileanu@upb.ro

beamforming or spatial filtering. For example, [1] presents a good synthesis over the methods used for sound localization and in [2] we can see that some of the current widely used high-end mobile devices use these techniques to improve reliability of speech recognition. Binaural hearing is an essential function, but we are able to distinguish between multiple speech sources even when we perceive them as coming from same direction. For example, during a radio talk show, and at some point it happens that all the guests speak simultaneously. If we stay concentrated, we can understand what each of the speakers is saying. This is a scenario where selective auditory attention (SAA) (or selective hearing – SH) shows its importance. This function is often referred to in technical literature as blind source separation (BSS).

BSS was treated in several studies during the last decade (e.g. [3], [5] and [6]). Their main majority makes use of microphone arrays but there are few approaches using a single microphone (e.g. [6]). In the mentioned papers BSS is treated as a signal processing problem but there is evidence that human selective hearing is a learned ability across more than 10 years (e.g. [12]) indicating that such simple techniques may not be the best choice. There are also studies that limit their scope to determining the number of competing speakers in a timeframe – [4], [7]. Their declared goal is to use this information as an input for future BSS algorithms. However they do not present data collected with the help of human listeners and do not operate on environment agnostic audio recordings as we attempt in this paper. We note that there are clinical studies targeting SAA. Paper [8] describes an experiment completed with the help of 44 child and adult volunteers that studies selective attention using nonlinguistic and linguistic probes. Study [9] presents the results of a research project on identifying the causes for SAA disorders. Even if the topic is of considerable importance, we could not find data that gives details about the actual performance of human SAA.

In this paper we describe the results of the SAA stress experiments that were made with the help of a group of native Romanian volunteers. We will focus on determining the maximum number of competing speakers that a person can follow. Our research goal is to create a method that would improve the results of BSS and as a first step we developed a “brute force” estimator based on dynamic time warping (DTW) computations in the time-frequency domain that tries to “guess” correctly the number of competing speakers. This method can be used as a trigger and tuner for future complex BSS algorithms. For example if we can detect the signal periods where there is only one active speaker we can extract his voice characteristics, or when there are multiple speakers we can detect their number and fine tune the BSS algorithms parameters. At this point labeling periods with multiple active speakers is done during voice activity detection phase (VAD) (e.g. [10], [11]). We used the same approach and the proposed estimator serves also as a high performance VAD.

Sections 2 will describe in detail the methodology used for the perception experiment, while section 3 will present the internals of the “brute force” estimator and the methodology used to determine its operating capabilities. Section 4 is dedicated to explaining the experimental results achieved with both the volunteers and the artificial methods and finally Section 5 and 6 present our conclusions and indicate future work steps.

2. Experiment Description

For the perception experiments we used a group of 31 male and female listeners aged between 21 and 37. According to [12], school aged children can be auditory selective but inflexible. The development of selective hearing continues during adolescence so at the age of 21, persons can present a well developed SAA. We selected volunteers that were within top 15% students considering Romanian grading system and demonstrated good concentration skills. In addition, the participants were motivated by making them compete.

2.1 Speech Sources

For creating the speech mixes, we selected a set of high quality audio samples produced by native Romanian speakers. 70% of the recordings are attributed to publicly known persons: famous actors, politicians, journalists, writers and business men. The rest of the 30% was generated by lesser known persons. A voice of a person can appear in more than one recording. We choose this approach in order to understand if the listener recognizes a known voice and if he is also able to learn and count new voices.

We anticipated that combining male and female voices in the same recording would ease the detection process and therefore only 20% of samples are produced by women voices and are added to the mix when the speaker count is higher.

As additional measures, we selected speeches that contain corporate language that is easily understandable by the audience, and also verified that there are no long pauses during the speech, so that after mixing, in the majority of timeframes all speakers are active. As probably deducted, each speech was recorded in its own environment therefore we cannot apply methods described in [3], [4] and [7] because we have no information about room acoustics.

2.2 Methodology description

We combined the speech files so that starting from second 0 of the recording until its end all speakers in the mix are active. Each listener had to follow 9 recordings, formed by mixing 2 up to 10 simultaneous speeches. The recordings were presented in a random order (to avoid the temptation to increment or decrement the previously given answer). After each track the listeners were asked the following questions:

- How many competing speakers did you hear talking?
- How did you identify each speaker?
- What made it difficult to identify the speakers?

As we will see in the results section, most of the answers can be grouped and we generated a statistic that can be used for further studies.

When analyzing the speech mixes, we realized that some speakers have stronger voices or they talk very loud so other speakers “get shadowed”. In order to compensate this issue, we used audio processing software to amplify the sounds attributed to speakers with softer voices and then we created the mixes again. When the recordings were played we used headphones setting an isolating but non-disturbing level.

Another important fact is that we measured the response time for all subjects in order to understand if there is a correlation between detection accuracy and listening time.

3. Spectrogram Distance Based Estimator

The idea of the method is simple and based on several responses we got from the volunteers. They stated that as the speaker count grew, the recording sounded more similar to noise. So in our approach we determine how different a multi speaker mix is from a sample produced by a single speaker.

We expect that as the number of competing speakers grows, we will be able to see a clear increase in the distance between a reference single speaker signal and a multi speaker recording. For the sets of experiments described in the current paper we will use multiple references (all the single speaker files used for creating the mixes) but in future studies we will investigate the possibilities of using a single reference produced by an unknown voice. In addition, future experiments will target a language other than Romanian. There is no reason that the current method won’t be applicable for other languages as it does not rely on language characteristics.

To continue with the method presentation, the similarity degree between the reference signal and the multi speaker recording can be extracted from a distance matrix described by (1). In this formula *D(i, j)* is also a matrix whose values represent the frame by frame *DTW* distances of signals *i* and *j*.

$$D(i,\, j) = DTW\,(SingleS_{S_{w_i}}, MixedS_{S_{w_j}}) \qquad (1)$$

Basically we take the spectrograms of the 2 recordings (single speaker and speech mixes) and compute the DTW distances between each window *i* and *j* obtaining a set of floating point values. We then derive a single value metric (similarity degree). At this stage we just add the elements of D matrix but we admit that this method can be tuned to use only an essential set.

We now remind that 70% of the voices presented to the listeners group were known a priori. This is why we use the first half of the single speech signals as references (training data) and the other half to create the mixes (test data). We can now make the statement that the system "knows" the involved voices but does not know how they are combined and what they are saying. Fig. 1 represents the graphic summary of this description:

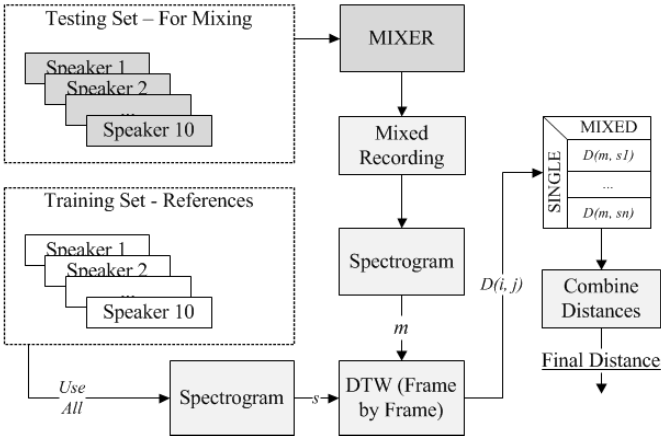

Fig. 1: Spectrogram Distance Based Estimator

The proposed method is designed to compare the entire mixed recording with all the single speaker references and estimate the cumulated number of speakers. Unfortunately this reduces the practical value of the estimator because it requires a long analysis time. In this study we try to determine the minimum speech duration necessary to estimate the number of competing speakers. We therefore selected a number of starting points in the speech samples and progressively increased the analysis window length until we were able to see a clear pattern of separation between distances.

Fig. 2 illustrates how we determined the minimum duration. It shows that as the analysis duration grows, the similarity between the single speaker reference and the test samples decreases (distance increases). When the target frame is too low, the "distance signals" intersect and therefore we need to increase the frame size. The point from where we can see a clear divergence between different "distance/similarity signals", can be considered as a lower limit of the analysis timeframe.

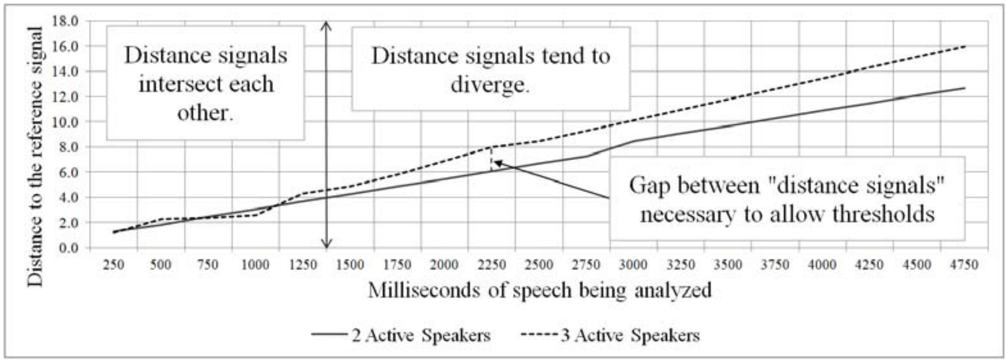

Fig. 2. Determining the minimum duration needed to estimate the number of speakers

The proposed procedures require a huge number of DTW instances so we invested some effort in optimizing the routine implementing it in a SIMD friendly fashion. By using a dedicated compiler we reduced the experiment execution time by a factor of 160X, on a 2.5 GHz quad core processor, which reduced our waiting time from months to hours.

4. Experimental results

In this chapter we will describe in detail the results achieved by human listeners correlated with the results achieved by using the artificial method described in section 3. We will also focus on determining the minimum duration needed for the estimator to give reliable answers, in order to assess its practical value.

4.1 Results obtained by human subjects

Perhaps the most interesting result is to see how many competing speakers can be detected by an adult person. In Fig. 3 each bar quantifies the percentage of listeners that counted incorrectly the number of active speakers noted on the x-axis. We can see that only 4% of the listeners gave wrong answers when presented a mix of 2 speeches but when the mix grew to 4 competing speeches, 52% were wrong.

When there were 5 competing speakers, less than a third of volunteers gave the correct answer. We can also see that the error grows with a logarithmic trend, clearly showing that higher counts of simultaneously active speakers can easily confuse a human listener. Each of the listeners was asked a set of 9 questions and we counted 87 correct answers from a total of 279. This means that the volunteers obtained a correct detection ratio for the number of competing speakers in a speech mix of 31%.

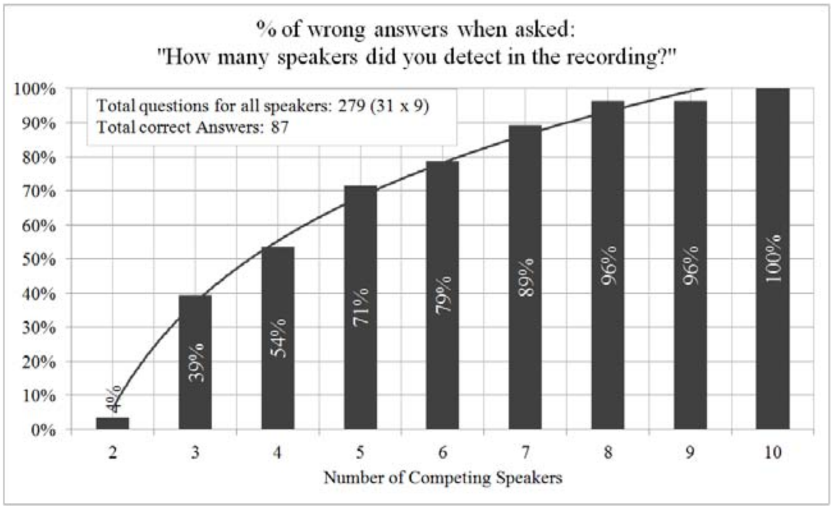

Fig. 3. Detection errors for competing speakers

Table 1 aggregates the observations of listeners with respect to the methods they used to detect the speakers.

Table 1

Reported detection methods

Observation / Method of detecting speakers	% of listeners used it
Learned voices from previous recordings	78%
Recognized a known voice	67%
Was able to follow transmitted information	53%
Recognized different genders	46%
Just guessed where speaker count was high	39%
Detected different speech paces	32%
Reported hearing other languages	17%
Used silence periods to identify new speakers	14%
Reported words that are repeating	10%

The results are very interesting as they give clues on how selective hearing or SAA works. From the results above we can see that the voice characteristic of the speaker plays a key role in SAA. In the same category we can include his gender. A surprisingly low percentage of volunteers reported that they could follow speech – only 53%. Most of responders claimed that they just acknowledged pieces of speeches but they could not follow coherently to a certain speaker. Data also shows that the speech pace and the silence periods were used as a factor in identifying new speakers.

In our experiment we also investigated if the listening time is able to reduce the detection error. As said, all speakers in a mix start at moment 0 and have a fluent, coherent speech on the same topic until the end of the recording. So basically there is no chance that one speaker starts later than other. Fig. 4 shows the correlation between listening time and speaker detection accuracy. The black bars represent the accuracy of detection for each listener and the gray bars figure the average number of seconds spent per each recording for the same person

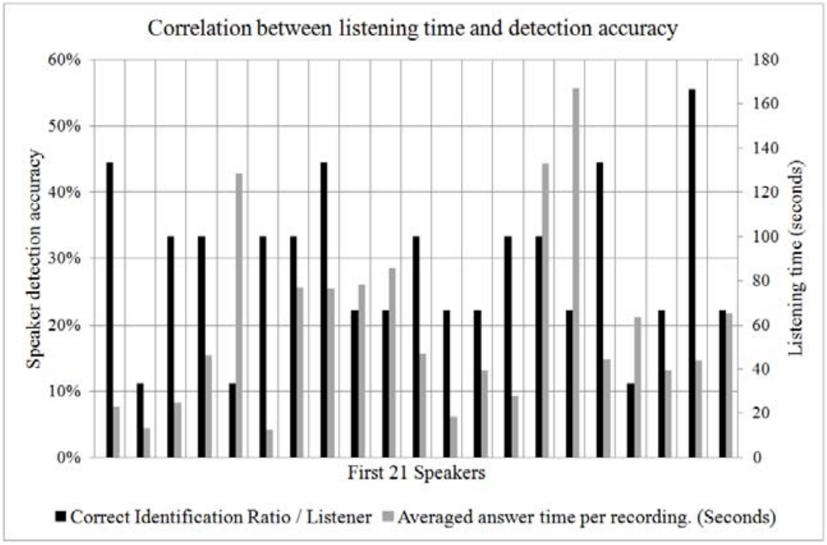

Fig. 4. Accuracy correlated with listening time

We can observe that listening time does not appear to drastically reduce the identification errors. For example we can see listeners that spent on average 20 seconds per recording but demonstrated accuracy equal to subjects that spent 130 seconds on average per recording. However this fact can have 2 potential conclusions: *Listening time does not necessarily influence* competing speakers detection ratio; *Selective hearing performance can differ a lot* from one individual to another and therefore the influence of time needs to be investigated further. In order to refine this test in the future, listeners could be asked to attend multiple sessions with different sets of mixes presented in each session – but with similar difficulty. A different response time constraint should be added in each session in order to generate a per user statistic with different response times per similar experiments.

4.2 Spectrogram distance based estimator results

Unlike a human subject, our estimator works based on a fixed algorithm presented in Chapter 3. Analyzing the performance of the method involves 2 steps:

- Analyzing the entire speech samples and estimating the cumulated number of speakers.
- Determining the minimum signal duration for which the estimator can generate a reliable output.

4.2.1 Determining the number of competing speakers in the recording

To produce Fig. 5, we considered each single speech sample as a reference (training set) and compared it with all the test speech mixes. We enforce that the training set is different from the test set.

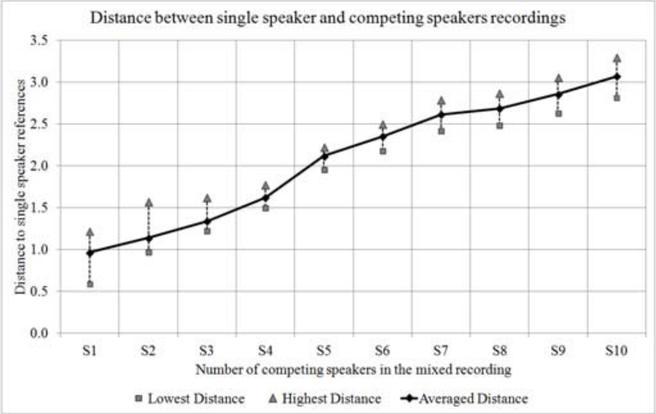

Fig. 5. Distance trend as number of competing speakers grows

In order to derive a single floating point value as the distance metric between 2 sound files we added the values in the D matrix defined in section 3. In this case, the metric (similarity degree) represents the sum of frame by frame distances between single speech and multi speech samples.

Fig. 5 shows that as the number of competing speakers in a recording grows, the signal becomes more different from a sample produced by a single speaker. The vertical dotted lines show the dispersion of the obtained distance, considering all the single voice references.

We can easily observe a linear growth trend for the averaged distance and the same trend is respected by the minimum and maximum distances. This means that the system will have increased chances (comparing with human listeners) of responding correctly even for higher speaker counts.

In order to extract the number of competing speakers detected correctly, we computed a linear regression on the curve that represents the averaged distances and then we defined a set of equally spaced intervals associated for each speaker count. Therefore the detected speaker count can be computed using (2):

$$N = \left\lfloor \frac{\sum_i \sum_j D(i, j)}{\left(\frac{MaxT - MinT}{10} \right)} \right\rfloor \tag{1}$$

Using (2) we determined optimal MaxT and MinT thresholds to get best detection results. In the end, the estimator based on distance computations between signal spectrograms achieved a correct identification ratio of 66%.

4.2.2 Determining the minimum frame duration for the estimator

In order for the estimator to have a practical value and to allow the possibility of integration in a state of the art LV-CSR, it must operate on frames small enough to permit a real time functioning. For example, if the estimator needs 1 minute of speech to determine the cumulated number of active speakers, it can be of no use because the real valuable information is when multiple speakers are active and how many.

We determined statistically the minimum frame duration needed for the estimator to give reliable results by using all single speaker files as references. We used all the mixed speech files to generate 150 recordings to be analyzed. We then applied the technique described in chapter 3, highlighted by Fig. 2 and determined the analysis duration necessary to show a clear divergence between “distance signals” associated to multi speaker recordings.

Fig. 6 reveals that as the minimum required distance between recordings (*that mix consecutive number of speakers*) grows so does the minimum analysis duration. We need that the distances between recordings are sufficiently high to allow labeling by using a set of thresholds. We set the minimum gap between 2 recordings to be higher than 40% of the length of a value interval associated to each speaker count. This interval is the difference between *MaxT* and *MinT* from equation (10), divided at the number of speakers. We chose this approach in order to ensure that the samples are separated enough to enforce good detection ratio after performing a linear regression. This decision can be described by equation (3), where D_{N+1} and D_N are the distances to the reference files of recordings mixing N+1 and N speakers.

In conclusion we established that for optimal results, the estimator needs at least 2.3 seconds of speech. However if the quality of captured sound is increased and speakers have very distinct voice features this threshold can decrease. The system can also function with less than 2 seconds of speech in its analysis buffer but this will lead to a more difficult method of labeling each frame of speech with the corresponding speaker count.

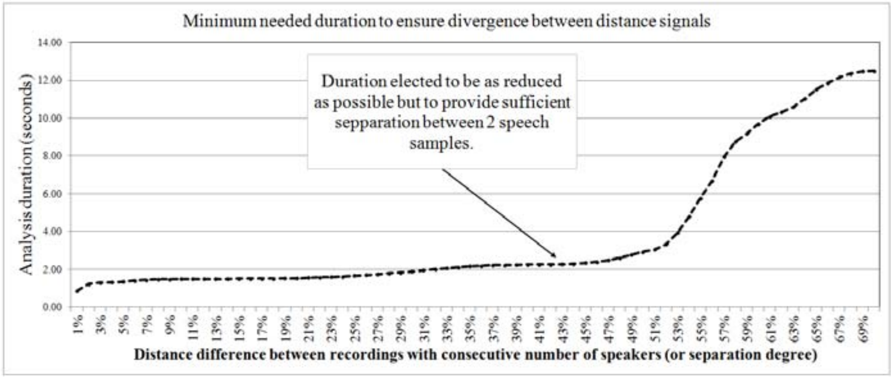

Fig. 6. Determining the minimum speech duration required by the estimator

$$D_{N+1} - D_N \geq 0.4 * \left(\frac{MaxT - MinT}{10} \right) \tag{3}$$

5. Conclusions

In this paper we described some of the characteristics and performances of human selective hearing with the help of a native Romanian volunteer group. The study can be naturally extended to other languages (e.g. Latin languages because they share the same family with Romanian) as it does not depend on language specific details.

Results show that human subjects have great difficulties in estimating the correct number of competing speakers in a recording with 4 or more. In order to estimate this number, the listeners claimed that they tried to identify voice features, different speech paces or pause periods. Less than 55% reported that they can actually focus on a single speaker and acknowledge the transmitted information – when the speaker count was high. We saw no correlation between the time spent for listening the recordings and the accuracy of the responses.

While the group of listeners estimated correctly the speaker count in 31% of the cases, we presented a simple estimator based on distance computations between signal spectrograms that achieved a correct detection ratio of 66%. However even if showing a lower accuracy, the human subjects were able to describe the voice characteristics of the detected speakers, unlike the automated method that just gives a single value as output. In addition we determined that the estimator needs about 2.3 seconds of speech to determine the number of competing speakers, making it a possible candidate for improving the performances of a LV-CSR operating in multi speaker environments.

6. Future work

As stated in the first section, our research goal is to contribute to the accuracy of BSS. Therefore we consider using the results of the proposed estimator for improving the performance of BSS methods. To achieve that, the system could be designed to auto tune thresholds based on the loudness and signal to noise ratio of the sound samples. This would improve robustness and reliability.

We will make further investigations to determine whether it is critical to use multiple single speaker references, or the accuracy drop won't be significant for a single reference. Nevertheless the proposed perception experiment could be realized on more volunteers and with more tests, like using known voices of friends or colleagues to perform the mixes, or creating a more complex scenario for determining the correlation between listening time and detection accuracy.

Acknowledgment

A part of this work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU /159/1.5/S/134398 and by the PN II Programme "Partnerships in priority areas" of MEN - UEFISCDI, through project no. 32/2014.

REFERENCES

1. *Wang, L and Brown G. J.:* “Computational Auditory Scene Analysis”, John Wiley & Sons, ISBN 0-471-45435-4, 2005
2. www.mactech.com: “Apple Patent Involves Audio BeamForming”, May 2010
3. *L. Hardesty.,* “Automatic speaker tracking in audio recordings”, MIT News Office, October 2013
4. *E. Zwyssig, S. Renals, M. Lincoln*, “Determining the number of speakers in a meeting using microphone array features”, Acoustics, Speech and Signal Processing (ICASSP), March 2012
5. *Ikeda, S. and Murata N.:* “An approach to blind source separation of speech signals”, Neurocomputing Vol. 41, pp1-24, October 2001
6. *Bach F. and Jordan M.:* “Blind one-microphone speech separation: A spectral learning approach”, 18th Annual Conference on Neural Information Processing Systems, 2004
7. *R. Kumara, R. Murty, B. Yegnanarayana*, “Determining number of speakers from multispeaker speech signals using excitation source information”, IEEE Signal Processing Letters, Vol. 14, No. 7, July 2007
8. *Coch, D., Sanders, L. D., Neville, H. J.:* “An event related potential study of selective auditory attention in children and adults”, Journal of Cognitive Neuroscience, Vol 17, Nr. 4, 2005
9. *Gomes, H., Duff, M., Ramos. M, Molholm, S, Foxe, J., Halperin, J.*: “Auditory selective attention and processing in children with attention deficit/hyperactivity disorder”, Journal of Clinical Neurophysiology, August 2011
10. *Lorenzo-Trueba, J.:* “Noise robust voice activity detection for multiple speakers”, International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2010
11. *Maraboina, S., Kolossa. D, Bora, P., Orglmeister, R.:* “Multi-Speaker voice activity detection using ICA and beampattern analysis”, 14th European Signal Processing Conference (EUSIPCO 2006)
12. *Werner, L.,* “Development of Auditory Behavior: Hearing Science”, Univ. of Washington Course, 2009