Detecting the number of competing speakers – human selective hearing versus spectrogram distance based estimator

Valentin Andrei, Horia Cucu, Andi Buzo, Corneliu Burileanu

University "Politehnica" of Bucharest, Romania

valentin.m.andrei@gmail.com, horia.cucu@upb.ro, andi.buzo@upb.ro, corneliu.burileanu@upb.ro

10.21437/Interspeech.2014-117

Abstract

This study describes an experiment designed to establish the maximum number of competing speakers that can be detected accurately by a human listener and compares the results with the ones produced by using a distance based estimator working in frequency domain. We mixed a set of high quality audio samples with continuous speech, produced by publicly known people (actors, journalists and politicians) and also unknown persons and then we played the tracks to each listener within a target group. The volunteers were asked how many cumulated speakers they counted and how they obtained the response. We observed that while human subjects showed a correct detection ratio of 31%, we were able to establish a set of equally spaced thresholds for the estimator in order to achieve 66% accuracy. The paper also summarizes the methods that were reported by the listeners to help in the detection.

Index Terms: selective auditory attention, voice activity detection, blind source separation

1. Introduction

Large vocabulary continuous speech recognition systems (LVCSR) 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 understand 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 correct speaker 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, 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 widely used high-end 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, at some point it happens that all the guests speak simultaneously. If we stay focused we can understand what each of the speakers is saying. This ability is called selective auditory attention (SAA) or selective hearing (SH). Its function is often referred to in science as blind source separation (BSS). Several studies on this topic were presented in the last decade and the main majority of methods rely on parameters that describe the geometry of the recording environment. As presented in [3] a time-frequency domain algorithm is adapted for a fixed virtual room, but no methods of auto tuning are presented. In [4] we can study a method of separating competing speech by segmenting the spectrogram in several disjoint sets. The paper does not give performance numbers, but it is one of the first methods trying to replicate the behavior of SAA without using environment geometry.

SAA is being studied from the medical point of view and the main goal of the research is to document occurring neurological events. In [5] we can read a study completed with the help of 44 child and adult volunteers that tries to study selective attention using nonlinguistic and linguistic probes. In [6] the results of a research project that aimed to identify the causes for SAA disorders are presented. Even if the topic is of considerable importance, we could not find data that gives details about the actual performance of human SAA.

The current article describes the results of the SAA stress experiments that were produced by a group of native Romanian speakers. We will focus on determining the maximum number of simultaneous speakers that a person can follow. From a neuroscience perspective, knowing this information could be a first step in developing a functional model for the SAA. Another important result of the study is that each person described how the active speakers were identified and we were able to generate an interesting statistic.

However, our research goal is to improve 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" the number of competing speakers. This method can be used as a trigger and tuner for future complex BSS algorithms. For example we can detect the signal periods where there is only one active speaker, extract his voice features, and when there are multiple speakers we can detect their number and fine tune the BSS algorithms. While we noted several studies – e.g. [7], [8] – that improve the results of voice activity detection (VAD) in multi speaker environments, none of them are able to determine the exact number of voices active in a timeframe.

Sections 2 and 3 will present in detail the perception experiment methodology and the "brute force" estimator. Section 4 is dedicated to explaining the experimental results and finally Section 5 presents our conclusions.

2. Experiment description

For our experiments we used a group of 31 male and female listeners aged between 21 and 37. According to [9], 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% of students considering Romanian grading system and demonstrated good concentration skills. In addition, the participants to this study were motivated by making them

compete on who has the better accuracy in detecting the right number of active speakers from each recording with the winner being rewarded.

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, making BSS algorithms (i.e. [3], [4]) ineffective.

2.2. Methodology description

We combined the speech files so that starting from moment 0 all speakers in the mix are active till the recording ends. There is no chance that some of the voices make their entrance later than the start of the recording.

Each listener had to follow 9 recordings, formed by mixing 2 up to 10 simultaneous speeches. The recordings were presented in the following order: 6, 8, 3, 7, 9, 10, 5, 2, 4 – where each of the numbers represent the number of active speakers in that audio sample. The idea of the sequence is just to be random (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 to question 2 can be grouped and so we generated a statistic that can be used for further studies. The last question is rather general, and was designed to get additional information unmentioned when answering the first 2 questions.

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 method is simple and based on several responses we got from the volunteers. They stated that as the speaker count grew, the recording sounded like noise. So we determine the distance between a speech mix and a single speaker reference. This metric is derived from a distance matrix described by (1).

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

Basically we take the spectrograms of the 2 recordings (single speaker and speech mix) and compute the DTW distances between all window pairs (***i, j***). The resulting *D* matrix can be processed to create a single value metric. At this stage we just add the elements but we admit that this method can be tuned and this will be treated in future work.

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 state that the system "knows" the involved voices but does not know how they are combined and what they are saying. Figure 1 represents the graphic summary of this description:

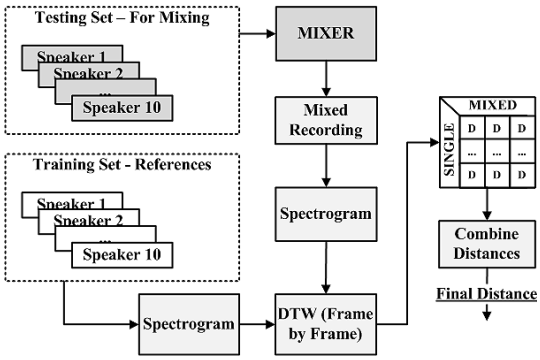

Figure 1: *Estimator Distance Generation*

The proposed procedure requires a huge number of DTW instances so we invested in optimizing the routine. We wrote a C library that was called from MatLab using a single instruction multiple data (SIMD) friendly implementation and a dedicated compiler for the target architecture. With this solution and combined with the fact that MatLab uses an interpreted language 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 detail the results achieved by human listeners correlated with the results obtained by using the artificial method described in section 3.

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 figure 2 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, 54% were wrong.

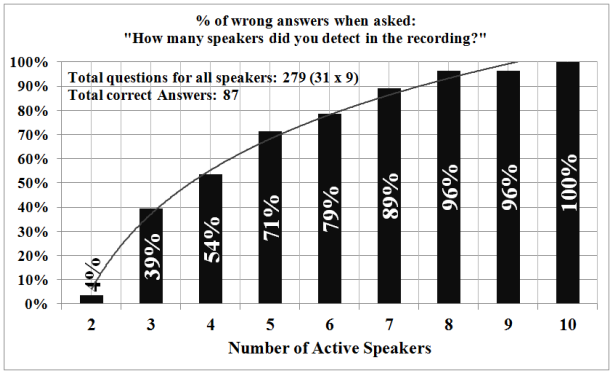

Figure 2: *Detection errors for competing speakers*

When there were 5 competing speakers, less than a third of volunteers gave the correct answer. We can also see that the error grows somewhat logarithmically towards 100% 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%.

After the interview we asked each listener to describe how he detected the speakers in the recordings. The answers are aggregated in table 1.

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 SAA works. It is important though, to admit that the results have a moderate confidence level as each volunteer is unique can express in his own way the listening experience.

From table 1 we can see that the voice characteristic of a 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 a certain speaker. Data also shows that the speech pace and the silence periods were used as a factor in identifying new speakers. For example, some listeners reported that they heard a man talking rapidly and part of them said that they could follow that man because it kept their attention focused. The fact that 17% of listeners claimed to hear other languages than Romanian reflects the difficulty of the task, especially if the speaker count is higher than 5.

In our experiment we also investigated if the listening time can 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.

Figure 3 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.

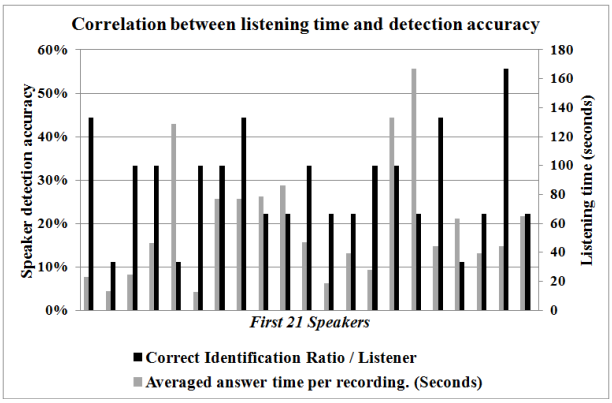

Figure 3: *Accuracy correlated with listening time*

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. It returns as output just the estimated speaker count and does not give other information. However the method can be modified to produce an output for each signal window. To produce figure 4, we considered each training single speaker file as a reference and compared it with all the test speech mixes. We enforce that we used different speech signals for references and for creating the mixes.

In order to obtain a single floating point value as the distance metric between 2 sound files we added the cells in the *D* matrix defined in section 3. This approach can be optimized by determining another schema of combining the values of *D* matrix in a single value. This will also have great impact on the speed of the method since it will greatly reduce the number of DTW computations.

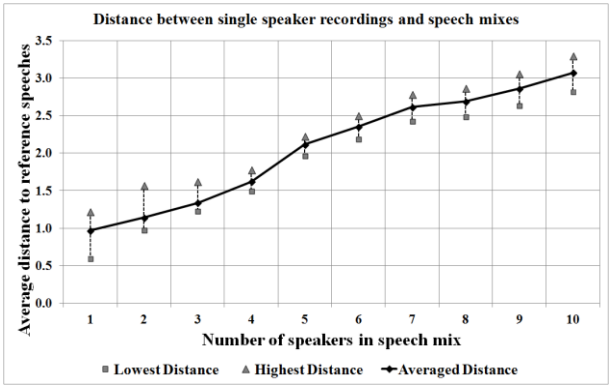

Figure 4: *Distance thresholds trend for each speaker count*

Figure 4 shows that as the number of competing speakers in a recording grows, the signal becomes more distant 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 by our system, we computed a linear regression on the curve that represents the averaged distances and defined a set of equally spaced intervals associated for each speaker count. Therefore the detected speaker count can be computed using (2):

$$N = floor\left(\frac{\sum_i \sum_j D(i,j)}{\left(\frac{MaxT - MinT}{10}\right)}\right) \quad (2)$$

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%.

5. Conclusions

In this paper we described some of the characteristics and performances of selective hearing for native Romanian speakers. The study can be naturally extended to other languages (starting with Latin languages as Italian, Spanish, French, etc.) 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.

6. Future work

As stated in the first section, our research scope 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 current algorithm will be modified to give a frame by frame estimation. Also, the system could be designed to auto tune thresholds based on the loudness of the sound samples. This would improve robustness and reliability.

In addition we will make further investigations to determine whether it is critical to have the speech samples produced by the targeted speakers or if any reference set of samples can generate similar results.

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.

7. Acknowledgements

The work has been partially 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 partially developed with the financial support of the Romanian-American Foundation. The opinions, findings and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect those of the Romanian-American Foundation.

8. 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] Ikeda, S. and Murata N., "An approach to blind source separation of speech signals", Neurocomputing Vol. 41, pp1-24, October 2001

[4] Bach F. and Jordan M., "Blind one-microphone speech separation: A spectral learning approach", 18th Annual Conference on Neural Information Processing Systems, 2004

[5] 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

[6] 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

[7] Lorenzo-Trueba, J., "Noise robust voice activity detection for multiple speakers", International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2010

[8] 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)

[9] Werner, L., "Development of Auditory Behavior: Hearing Science", University of Washington Course, 2009