Counting competing speakers in a timeframe – human versus computer

Valentin Andrei, Horia Cucu, Andi Buzo, Corneliu Burileanu

Speech and Dialogue Research Laboratory, University "Politehnica" of Bucharest, Romania

am_valentin@yahoo.com, horia.cucu@upb.ro, andi.buzo@upb.ro, corneliu.burileanu@upb.ro

Abstract

We propose an automated solution for computing the number of simultaneous active speakers within a timeframe. The method is studied in parallel with a perception experiment realized with the help of 28 volunteers that were asked to detect how many speakers talk simultaneously in several recordings with variable length. For this study we focus on how listening time and the usage of familiar voices in the recordings impact the correct detection ratio. Regarding the automated method we discuss the influence of noise and the evolution of detection error determined by the speech duration. We observe that when capturing clean speech sources, the method is 76% accurate even for 10 simultaneous speakers, considering speech lengths longer than 3.5 seconds. The volunteers did not systematically detect correctly more than 4 competing speakers even when listening up to 80 seconds.

Index Terms: competing speech, auditory scene analysis, speech perception, voice activity detection.

10.21437/Interspeech.2015-673

1. Introduction

Auditory scene analysis (ASAN) is a human reflex ability allowing us to distinguish multiple speech sources in a competing speaker environment, and focus our attention towards the speaker of interest (selective auditory attention – SAA). The challenge of developing computer based methods for ASAN has been accepted by a fair number of researchers and can yield useful applications in contexts not limited to ambient assisted living, forensics, blind source separation, smart surveillance and automated data mining. Taking decisions using ASAN related methods is a difficult task due to the complexity of the mixed speech signals. In addition there is no sufficient understanding on how human SAA works; in order to develop biology inspired methods.

A key information extracted using computer based ASAN is the number of speakers producing speech at a certain moment in time. This process can follow the voice activity detection phase and add a new functionality: the ability to output the active speakers count per analysis frame. In [2], monaural and microphone array techniques are combined to produce the voice activity masks, augmented with competing speaker count. The method can work accurately for use cases where speech sources are not far from the sensors and are not close to each other. In [3] the authors study single channel ASAN techniques and use a Bayesian methodology for speech source separation. The algorithm is an improvement over referenced approaches but without being aided by the speaker count information, it is yet far from being adopted widely. Looking more to single channel speech separation, we observe that deep belief neural networks are studied as a potential instrument for ASAN, e.g. in [4]. Focusing back to detecting the number of competing speakers, we highlight studies that use visual information like detecting the speakers' lips movement (e.g. in [5], [6]) to estimate the number of active sources. In [5] there is strong evidence that knowing the number of active speakers in the analysis window provides critical information to source separation algorithms. However, visual context is hard to capture, lowering the method's potential of large scale adoption.

In previous work [1], a methodology for evaluating the limits of human SAA was presented. Studies [8] and [9] fall within the same research area by analyzing SAA development for children. However, they are focusing on medical aspects, like early disorder detection. In [1], by using 31 listeners, we were able to detect how many competing speakers can be accurately detected by a person in a single channel recording. The present paper adds value by re-iterating the experiment with improved methodology and by trying to estimate the impact of using well known voices (e.g. co-workers' voices) over the accuracy of the detection. In addition, recordings quality and listening time are strictly controlled in the new set of experiments. The perception study was driven by a custom software application that we developed in order to ensure exact same conditions for all participants.

For comparison with the results achieved by human listeners, in [1] we proposed an automated solution for detecting the number of active speakers. Our experiment used single channel samples because the goal was to compare the output of the algorithms with the results of human listeners not being aided by binaural hearing. The current study extends the analysis and presents the minimum time slice (listening time) for the computer routine to take a decision. This is an essential step forward for integrating the method into a more complex system, targeting for example speech blind source separation. The influence of noise over the proposed algorithm is discussed in detail for studying the adoption potential.

Sections 2 and 3 will present the perception study new methodology as well as enhancements to the automated solution. Section 4 will detail the results achieved by human listeners in parallel with the performances of the automated method working on clean and also noisy speech samples.

2. Perception study design

A number of 38 persons participated to this perception study. We divided the volunteers into 2 groups: A and B. Participants from Group A and B have never met before. All members from Group A know each other because they are co-workers. Group A had 19 volunteers: 10 of them generated the test speech recordings while the other 9 had listeners' roles. Group B gathered 19 persons that were only listeners, so the total number of listeners was 28.

A few observations about the recording phase:

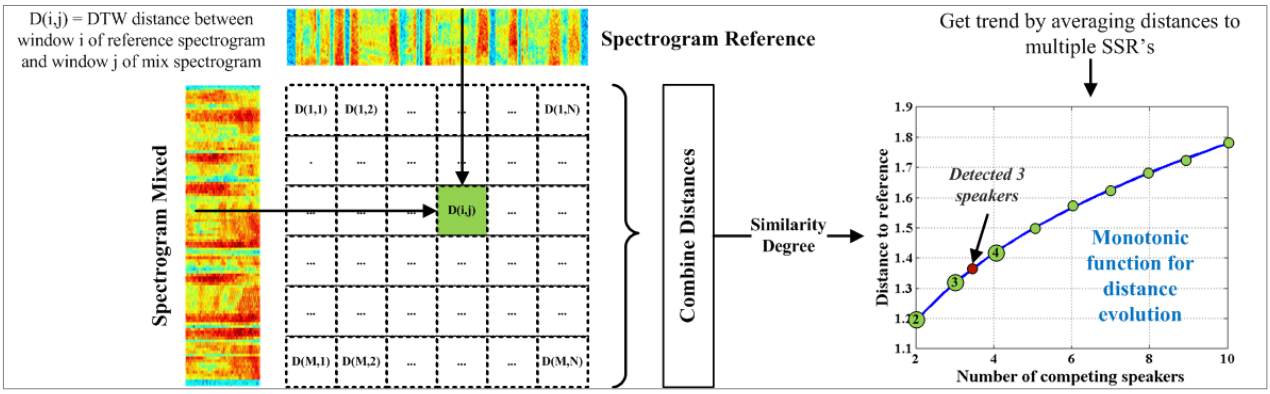

Figure 1: *Competing speaker counting solution – similarity degree and count mapping*

- The recordings were taken using a high quality recording device, in a soundproof room.
- Every speaker sat in the exact same place in the room.
- We used only male speakers because speakers of different genders are easier to differentiate. In addition, each speaker had to read from a different topic (e.g. chemistry, biology, physics, computer science, etc.).
- The recordings were generated by group A members that did not participate as listeners. We tried to observe the impact of known voices over the detection accuracy.

The collected speeches were mixed into 9 single channel recordings with speaker counts ranging from 2 to 10. It's important to highlight that all the speakers in the recording are active from the beginning to its end. The method of combining speakers is not incremental (e.g. the mix with 4 speakers does not include all the speakers that contributed to the one with 3). Each mix has random speakers. Also, before producing the mixes a fading step on some single speaker recordings was ran. This was because some speakers spoke louder, shadowing others, and we wanted to avoid this effect.

The 28 listeners are aged between 22 and 32 and they were asked to determine the number of competing speakers per recording. They detailed if they recognized a known voice, if they were able to follow a certain speaker or detect the speech topic. Compared with [1], several improvements to the methodology were made:

- We developed a software application with a strictly controlled flow, where the participant could not stop the current recording, even if he knew the answer. This was done in order to force a minimum listening time per recording and to ensure identical conditions for listeners.
- For each competing speaker count, 5 different recordings of 5, 10, 20, 40 and 80 seconds were played. For every targeted speech length a mix of different speakers was used.
- The application played the recordings in a totally random order with respect to both competing speaker number as well as recording duration.

3. Automated method description

The purpose of the automated method is to determine how different a recording that has multiple active speakers is from a set of single speaker references (SSR). We rely on the fact that as the speaker count in the recording grows the difference compared to a SSR will also grow following a trend that can be modeled using a monotonic function. If the trend can be modeled a set of thresholds can be determined on the trend's curve in order to determine the speaker count. We also expect that for higher speaker counts the error will grow, as in the case of human listeners.

3.1. Similarity degree and speaker count mapping

Figure 1 reflects the entire process of determining the similarity degree between a simultaneous multi-speaker recording and a SSR and the mapping to a speaker count.

The first step is to compute a spectrogram for each of the signals. We used a window length of 0.5 seconds with an overlap of 0.125 seconds and 1024 points for the FFT. The sampling frequency was set to 11025 Hz. We tried other feature extraction methods like PLP and MFCC but we discovered that this approaches generate a higher error rate. The reason is that MFCC and PLP are efficient when using narrower signal frames (≤ 0.1 seconds) while our method uses larger windows to better capture the effects generated by competing speakers. The selected window and overlap length were observed to determine performance sweet spots, after running multiple automated experiments.

The next phase is comparing all possible pairs of window spectra computed for the two recordings using dynamic time warping (DTW). For example, when using 2 recordings of 10 seconds generating 26 windows each, 676 DTW distances are computed in a 2D structure. In order to obtain a single similarity value, we add all the values in the distance matrix. We discovered that this approach generates good accuracy. A method of eliminating distances in the 2D structure would however decrease the computation time of the solution.

The final step is correlating the similarity degree with a previously computed trend. We observed that the trend can be modeled by a monotonic function (in most cases, if high quality SSR are used, the trend is linear). For each speaker count we can compute a set of thresholds on the trend. The number of active speakers results from the closest threshold lower than the distance to the SSR.

The trend and thresholds are obtained statistically, before runtime, by considering multiple SSR and averaging the similarity degrees. If the number of used SSR is higher, the extrapolation capability of the solution is expected to increase. In the case of signals heavily affected by noise, it is possible to compute the distance trend and the thresholds dynamically, during runtime, for environmental adaptation.

3.2. Reference selection

Multiple SSR are needed to determine statistically the distance increase trends and the thresholds associated with different speaker counts. Theoretically, on a fixed dataset a single reference is enough to highlight the trend as the active speakers count increases. However, selecting more SSR will improve the extrapolation capabilities of the solution.

In contrast with the methodology proposed in [1], for the current experiment we used as SSR a set of recordings that were not involved in the mixing process. The SSR are produced by 5 members of group A that also had listener roles. We did not use any of the SSR in the mixing process. The experiments were done on a set of 5 SSR used to create the distance trend and thresholds discussed in the previous section.

4. Experimental results

In this section we present the results obtained for the perception experiment as well as for the automated solution. For the perception study we are interested in determining if using familiar voices in the recordings determines a change in detection error. Also the effect of listening time is highlighted.

Regarding the automated method we are interested in determining the dependency between time frame length and detection error. We also discuss the impact of noise over the detection accuracy.

4.1. Perception study

Figure 2, shows the averaged declared speaker count for each multi-speaker mix alongside with the real speaker count. We can observe that while the real speaker count follows a linear path, the declared speaker count trend is more likely asymptotic. This shows that for higher numbers of simultaneous active speakers, human listeners always underestimate the real number. For lower counts, they tend to overestimate. We also highlight that for mixes with more than 5 speakers, the estimation error grows reflecting that the volunteers did not systematically detect correctly more than 4. This finding is in line with data published in [1].

Another remark is that time appears to slightly improve detection accuracy but this effect is subtle and in order to see a clear impact, many more recording sessions and volunteers are needed. In figure 2, we can see that the dotted line corresponding to 80 seconds of listening is the closest to the real speaker line. We also can observe that the dotted lines corresponding to 5 and 10 seconds of listening are on average the most distant compared with the real speakers' references. The measurements reveal a 27% detection error reduction going from 5 seconds to 80 seconds of listening.

In order to determine the effect of known voices over the detection accuracy, we define the *Absolute Detection Error Rate* (ADER) metric, associated for each speaker count and described by equation (1). N represents the number of volunteers, K represents the real number of active speakers and C is the declared speaker count. An ADER of 0.05 for a mix of 2 speakers, means that 1 listener out of 10 missed the correct speaker count by one.

$$ADER_K = \frac{\sum_{i=1}^{N} |C_i - K|}{N \cdot K} \quad (1)$$

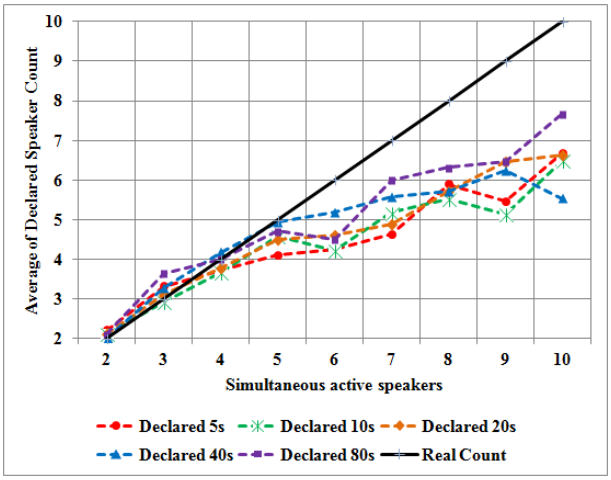

Figure 2: *Detection errors of competing speakers*

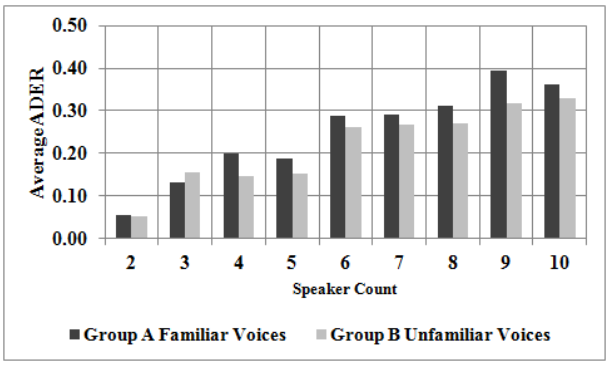

Figure 3: *Impact of familiar voices*

In figure 3 we plotted the average ADER for each speaker count, considering the 5 variable duration recordings. The most interesting finding is that even if the mixes are created using familiar voices that does not improve the ADER. On the contrary we measured an average increase in detection error of 13%. Except for 3 simultaneous speakers, in all cases the detection accuracy was lower when familiar voices were used in the recordings.

The justification of this fact lies in the comments given by group A's listeners. Their main majority stated that when they were able to recognize a known voice, they continued following the detected speaker, "filtering" out all the other speakers. This can be viewed as a normal effect of human selective auditory attention. Group B listeners were not able to recognize voices therefore they were more focused on detecting as many speakers as possible. So the lower ADER achieved by group A is explained as follows: even if group A has the advantage of knowing the voices in the recording, the listeners have the disadvantage of being distracted by recognized voices, not focusing on detecting other speakers, as the members of group B do.

4.2. Noise and automated method accuracy

We use DTW as a method of comparing spectra associated to signal windows belonging to SSR and speech mixes. In the presence of noise, the similarity degree between a mix and a SSR will be affected systematically. The following paragraphs highlight some of the noise induced effects.

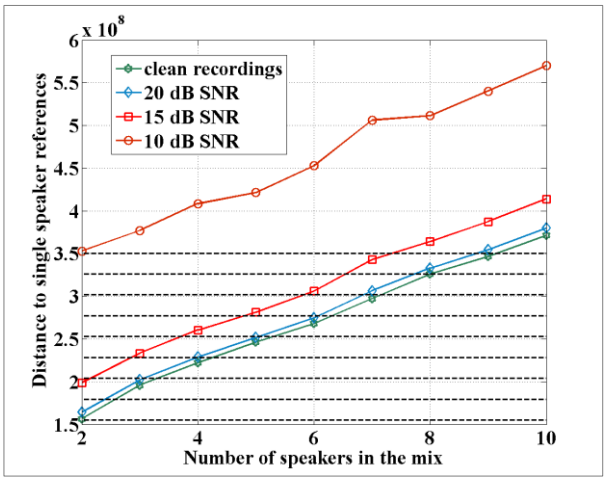

Figure 4: *Influence of noise over automated detection*

To obtain the data in figure 4, we used the entire length of each recording, which is 150 seconds. We then added various levels of white Gaussian noise to the speech mixes. The straight dotted lines represent the distances thresholds discussed in section 3.1. The clean mixes were compared to all the single speaker references and by averaging the similarity degrees, we obtained the distance thresholds. The colored lines represent the similarity degrees between speech mixes (affected by different noise levels) and the references. If a computed distance was greater than exactly N thresholds it means that the speech mix had exactly N+1 active speakers.

We can observe that the lowest green line, representing clean recordings generates the best accuracy. We see that all the distances generated by a mix with N speakers are greater than exactly N-1 thresholds. This shows that in the case of clean samples, the method has 100% accuracy, if we are considering 150 seconds of analysis duration.

On the opposite pole, we have the mixes with 10 dB SNR where all the distances are greater than all the thresholds. This reflects that the method had 0% accuracy. Looking at the mixes where SNR was 15 dB, every distance generated by N active speakers was greater than exactly N thresholds, meaning that the accuracy was also 0%. However, starting from 20 dB SNR, the results approach to the ones generated by clean samples, reaching 89%.

So we can state that if the distance thresholds are computed using clean recordings, the method gets 89% accuracy up to 20 dB SNR. However in the case of SNR lower than 20 dB, we can also observe a monotonic distance trend therefore a new set of thresholds can be computed, by using references or samples that are affected by the same noise.

4.3. Automated method operating time slice

In the previous section we discussed about the performance of the automated speaker counting method, considering 150 seconds of continuous speech. However this duration is prohibitive for adoption of the method into next generation speech recognition engines.

In figure 5 we can observe the effect of speech duration over accuracy and also notice the slight decrease of ADER when using higher SNR test recordings. The ADER decreasing trend, determined by longer speech durations is clearly observable.

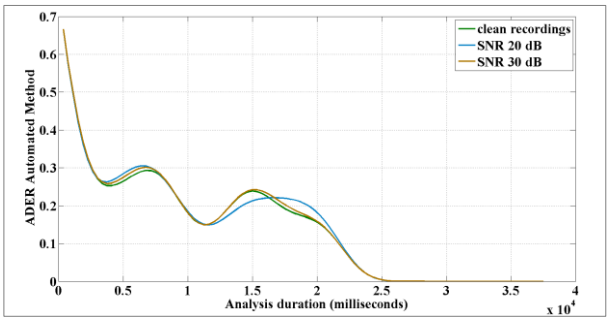

Figure 5: *Error depending on speech duration*

We observe that starting from 25 seconds recordings, the automated method is able to recognize up to 10 speakers in 100% of cases. However even 25 seconds is a long duration so we can select 3.5 seconds as the length of the analysis frame and still keep 76% accuracy. We must highlight that the detection errors for the automated method are more frequent considering higher speaker counts (≥7). So the 76% accuracy is sufficient for counting correctly up to 6 active speakers.

However speech length equal to 3.5 seconds does not necessarily mean that the algorithm lasts exactly 3.5 seconds. This depends on the number of SSR used and of the hardware capabilities of the device. We used a highly optimized DTW and a single x86 core running at 2.5 GHz in our tests and 3.5 seconds of speech mix were processed in roughly 2 seconds and we can state that the solution can be executed in real time.

5. Conclusions

In this paper we discussed in parallel the ability of human subjects to detect the number of simultaneous active speakers in a single channel recording and the design of an automated solution that performs the same task. We discovered that while the listeners were not able to detect accurately more than 4 speakers, even for 80 seconds of listening time, the automated method achieved minimum 76% accuracy for a maximum number of 10 competing speakers.

Using familiar voices is shown to increase the listeners' detection error with 13% compared to the case where speakers are unknown. This is because for high speaker counts, being able to follow a certain speaker has an impact over SAA efficiency. We observed that listening time decreased the detection errors produced by the volunteers with up to 27%.

The general justification for the high performance of the automated solution is the computational complexity. The speech length has higher impact on performance compared with human subjects. When using clean single speaker references, 3.5 seconds of speech analysis are enough for 76% accuracy while 25 seconds can increase accuracy to 100%. The solution works accurate up to 20 dB SNR. For lower SNR, an adaptive method for threshold computations can be designed due to the monotonic trend of the similarity to SSR.

6. Acknowledgments

This work was supported in part by the PN II Programme "Partnerships in priority areas" of MEN - UEFISCDI, through project no. 25/2014 and in part 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.

7. References

[1] V. Andrei, H. Cucu, A. Buzo, C. Burileanu, "Detecting the number of competing speakers – human selective hearing versus spectrogram distance based estimator," *Proceedings of 15th Annual Conference of the International Speech Communication Association*, pp. 467 – 470, 2014.

[2] J. Lorenzo-Trueba and N. Hamada, "Noise robust voice activity detection for multiple speakers," *Procedings of IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 1 – 4, 2010.

[3] S. Kammi and M. Reza Karami, "A Bayesian approach for single channel speech separation," *International Journal of Machine Learning and Computing*, vol. 2, no. 1, pp. 24 – 29, 2012.

[4] C. Weng, D. Yu, M. L. Seltzer, J. Droppo, "Single channel mixed speech recognition using deep neural networks," *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 5623 – 5636, 2014.

[5] B. Rivet, L. Girin, C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Communication*, vol. 49, pp. 667 – 677, 2007.

[6] V. P. Minotto, C. R. Jung, B. Lee, "Simultaneous speaker voice activity detection and localization using mid-fusion of SVM and HMM", *IEEE Transactions on Multimedia*, vol. 16, pp. 1032 – 1044, 2014.

[7] S. Akram, J. Z. Simon, S. Shamma, B. Babadi, "A state-space model for decoding auditory attention modulation from MEG in a competing speaker environment," *Advances in Neural Information Processing Systems*, vol. 27, pp. 460 – 468, 2014.

[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, no. 4, pp. 605 – 622, 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", vol. 123, no. 2, *Journal of Clinical Neurophysiology*, pp. 293 – 302, 2011.