MULTIMODAL SPEECH SEGMENTATION USING GAZE DATA AND SPECTROGRAM IMAGE FEATURES

Arif Khan[1,2], *Ingmar Steiner*[1,3]

[1]*Multimodal Computing and Interaction, Saarland University, Germany,*
[2]*Saarbrücken Graduate School of Computer Science, Saarland University, Germany,*
[3]*audEERING GmbH, Gilching, Germany*
arifkhan@coli.uni-saarland.de

Abstract: Nearly all automatic speech segmentation approaches rely solely on acoustic features, which differs from the way humans segment speech using phonetic annotation software.

In order to get closer to human-level precision in speech segmentation, we adopt a multimodal approach to improve the segmentation accuracy. To this end, we analyze a database of segmentation behavior collected using an eye tracker, obtained from human experts performing a manual segmentation task. This allows us to introduce gaze as an additional modality for automatic segmentation by transforming it into features for image based phoneme segmentation (ISeg).

Experiments were conducted for automatic speech segmentation, comparing the image-only, ISeg technique, as well as ISeg combined with hidden Markov model (HMM) based acoustic segmentation, with respective segmentation approaches conditioned on the gaze data. The results show that enhancing the image based segmentation with gaze information improves the accuracy of ISeg, as well as ISeg combined with HMMs.

1 Introduction

Phonetic segmentation is the process of inserting boundaries into the time domain of a speech signal, to match distinct phonetic units (*phones*), typically also labeling each unit. The output of phonetic segmentation is a set of boundaries, representing the start and end times of the phones of an utterance. Segmented speech data is an essential requirement for most speech-related applications and research. For instance, it is used in phonetic analysis, in text-to-speech synthesis systems, and for bootstrapping the training of acoustic models for speech recognition.

There are two main ways of obtaining phonetic segmentation, viz., manual and automatic segmentation. During manual segmentation, phonetic experts visually inspect short spans of speech and place boundaries within each span. In order to decide on a boundary, several sources of information are typically used: (a) the spectrogram as a time-frequency representation of the acoustic signal to detect boundaries based on the intensity changes in the frequency domain; (b) the oscillogram as a time-intensity representation; and (c) playback of audio segments to perceptually verify that a boundary suitably separates two phones. During the entire process, the expert uses a graphical user interface (GUI) and relies on prior knowledge to segment speech.

Manual segmentation tends to yield better quality than automatic segmentation [1]. However, it is very time-consuming, and expensive for large corpora; moreover, consistency can vary significantly across annotators (or even within the same annotator over time). As a solution, automatic segmentation is desired, as it is fast and reproducible. In automatic segmentation, acoustic features (commonly, mel-frequency cepstral coefficients (MFCCs) [2] or perceptual

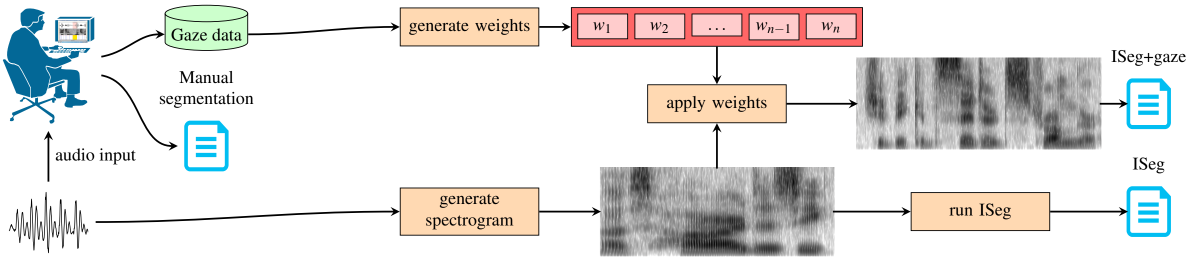

Figure 1 – (left) Data recording and (right) segmentation

linear prediction (PLP) [3] features) are extracted from the speech signal, which represent the continuous signal as discrete frames. These features are then used to train a model for segmentation. The use of acoustic features gives an acceptable segmentation quality for clean, well-recorded speech; however, these features are derived from only one of several information sources (acoustics) that humans use to segment speech. Therefore, we hypothesize that the quality of automatic segmentation can be improved by using more information sources in the segmentation process, such as a modality corresponding to the visual domain.

In order to find a candidate modality, we recorded human segmentation behavior to investigate the information sources they used to segment speech [4]. After analyzing the recorded data, it was found that most visual attention was focused on the *spectrogram* region of the Praat GUI and that gaze can serve as a correlate of annotator attention during the segmentation process [4]. This data is used to derive gaze features and used in the segmentation process to improve the quality. In Fig. 1, the data recording (left) and the feature processing step, i.e. how gaze data is transformed into weights to be used in segmentation, are shown. This makes the modeling of the segmentation more similar to the behavior of human annotators, who use more information sources than acoustics to segment speech.

The rest of this paper is organized as follows. Section 2 provides an overview of previous work on visual attention and spectrogram based speech segmentation. In Section 3, we briefly describe the gaze data of phoneticians and its processing. In Section 4, the combination of gaze data with other segmentation techniques – image based phoneme segmentation (ISeg) and hidden Markov models (HMMs) – is shown along with results. Finally, we conclude the paper in Section 5 and discuss potential limitations of the approach.

2 Background

Gaze is a strong indicator for attention [5]. When humans look at a scene, they continuously make eye movements called *saccades*. Between saccades, the eyes remain constant, resulting in *fixations*, whose duration is usually 200 ms to 300 ms [6]. New information is only perceived during fixations, as the saccades are rapid movements and the scene is blurred for perception [7]. There is some disagreement between researchers on whether to include saccades with fixations when computing gaze durations [6]; in this work, we have considered only fixations for computing gaze durations.

The relationship of fixation durations and the important regions of a scene has been established by many researchers. For instance, Loftus and Mackworth [8] showed that important objects in the scenes are fixated more and longer than less important objects. Kundel et al. [9] have reported that fixation durations are directly correlated with high informativeness in the scene.

The visual representation of speech, particularly in the form of *spectrograms*, has long been used by humans for labeling speech recordings [10]. Dennis et al. [11] have used spectrogram image features for the classification of sounds. Other researchers have already used spectrograms as a modality, along with the acoustic features, for automatic speech segmentation. In

Table 1 – The results of fusing ISeg [12] with HMM segmentation using the TIMIT corpus [13].

Method	CDR(%)	OS(%)	FA(%)	F-score	R-value
ISeg	78.07	7.74	27.54	0.75	0.78
HMM	84.57	7.73	21.50	0.81	0.83
ISeg+HMM (raw)	94.90	104.24	53.53	0.62	0.09
ISeg+HMM (pruned)	89.75	33.98	33.01	0.77	0.67

particular, Leow et al. [12] used spectrogram images as features for improving the accuracy of automatic segmentation. In their work, the spectrogram is treated as an image, and then image based techniques are applied to enhance the spectrogram. After this enhancement, the boundaries are computed using the ISeg algorithm. This approach is based on the fact that locations in the spectrogram where sudden changes occur in intensity along the time axis, are likely the candidates for phone boundaries. The results of ISeg are then merged with a conventional, HMM based segmentation.

We choose the ISeg algorithm to be used in our work because it directly uses the spectrogram for segmentation without any prior knowledge of the speech. Additionally, the results of ISeg are comparable to other unsupervised segmentation algorithms [14, 15, 16]. Table 1 shows the segmentation results of the ISeg algorithm and combining it with HMM based segmentation using the TIMIT corpus. The F-score and R-value, which were missing in the original paper, are also computed.

3 Segmentation data

In order to observe the human segmentation behavior, we used a multimodal corpus recorded from phoneticians performing manual segmentation tasks [4]. This corpus contains eye-tracking data from seven phoneticians (identified as “vp02” to “vp08”) with varying amounts of segmentation experience. Each participant was asked to segment the same 46 s speech recording using Praat [17]. The audio recording contains the standard passage, “The North Wind and the Sun” [18], read by a male native speaker of English. At the end of each session, the manual segmentation produced by the participant was saved.

Manual segmentation is subjective and experts do not always agree on individual boundaries. Even if the number of boundaries matches for two annotators, the exact timestamps will still differ to some degree [1, 19]. In order to check the inter-annotator agreement among the participants, we calculated the Fleiss’ kappa [20], which resulted in a score of 0.71, indicating a substantial agreement according to Landis and Koch [21].

4 Experiments and Discussion

We conducted several experiments in order to combine different segmentation results and investigate the effectiveness of gaze data for automatic segmentation. First, we explain the different types of segmentation used, and how they were obtained.

Automatic segmentation To get automatic segmentation for the audio, the WebMAUS forced aligner [22] was used, which uses an HMM based model to get the segmentation.

Manual segmentation The segmentation obtained at the end of each participant recording session, this serves as the ground truth for comparing the results with the HMM segmentation.

ISeg segmentation The recorded audio was segmented by applying the ISeg algorithm to the spectrograms, without using any transcriptions. The parameters used for computing the spectrogram are the same as in Leow et al. [12].

Table 2 – Explanation of performance metrics. N_T is the total number of boundaries found by the segmentation algorithm. N_H is the number of correctly detected boundaries, after comparison with the manual segmentation. N_R is the total number of boundaries found in the reference segmentation.

Measure	Formula	Description
Correct detection rate	CDR = $\frac{N_H}{N_R}$	CDR shows how many boundaries are correctly detected out of the total correct boundaries in the reference.[a]
Over-segmentation	OS = $\frac{N_T}{N_R} - 1$	OS is an indicator of how many additional boundaries were detected.
False alarm	FA = $1 - \frac{N_H}{N_T}$	FA shows the number of falsely detected boundaries.

[a] according to each participant's manual segmentation

Table 3 – The correct detection rate (CDR), over-segmentation (OS), false alarm (FA) rate, F-score, and R-value given as mean (μ) and standard deviation (σ) across all participants, for the different conditions of the segmentation experiment.

	CDR (%)		OS (%)		FA (%)		F-score		R-value	
Condition	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
ISeg	70.20	1.97	0.55	8.95	29.82	4.94	0.70	0.02	0.74	0.03
ISeg+gaze	78.72	3.17	34.56	8.75	41.30	4.35	0.67	0.04	0.60	0.07
HMM	67.53	5.01	−9.37	8.06	25.37	2.89	0.71	0.03	0.75	0.02
ISeg+HMM	71.10	1.92	3.48	9.20	30.94	4.85	0.70	0.02	0.73	0.03
ISeg+HMM+gaze	79.58	2.95	37.78	9.49	42.03	4.15	0.67	0.03	0.58	0.07

4.1 Performance metrics

The boundaries of the automatic segmentation are compared with the manual segmentation, which is taken as the reference. We used the evaluation criteria as given by Leow et al. [12] and Estevan et al. [16], and as explained in Table 2. A threshold of 20 ms is used for comparing a reference and segmented boundary.

No single metric defined in Table 2 alone penalizes the performance of the segmentation with respect to falsely inserted boundaries. Therefore, two additional metrics were used, the F-score and the R-value [23]. Both F-score and R-value range from 0 to 1, with a value of 1 being the ideal segmentation where all the segmented boundaries lie within 20 ms of the reference boundaries and where no additional boundaries were inserted. The F-score, (1), shows the trade-off between precision and recall and is defined as:

$$F = \frac{2.0 \times PRC \times CDR}{PRC + CDR} \tag{1}$$

where precision is given as $PRC = 1 - FA$.

The R-value shows the performance of segmentation against the increase in falsely inserted boundaries and is given as:

$$R = 1 - \frac{|r_1| + |r_2|}{200} \tag{2}$$

where r_1 and r_2 are defined as:

$$r_1 = \sqrt{(100 - CDR)^2 + (OS)^2} \tag{3}$$

$$r_2 = \frac{-OS + CDR - 100}{\sqrt{2}} \tag{4}$$

4.2 Fusing gaze data with spectrograms

An initial step in the ISeg algorithm is to convert the audio signal into a spectrogram S having N rows and M columns. To combine the gaze data with the image based segmentation, it should be converted into frames such that each column of the spectrogram corresponds to a single frame

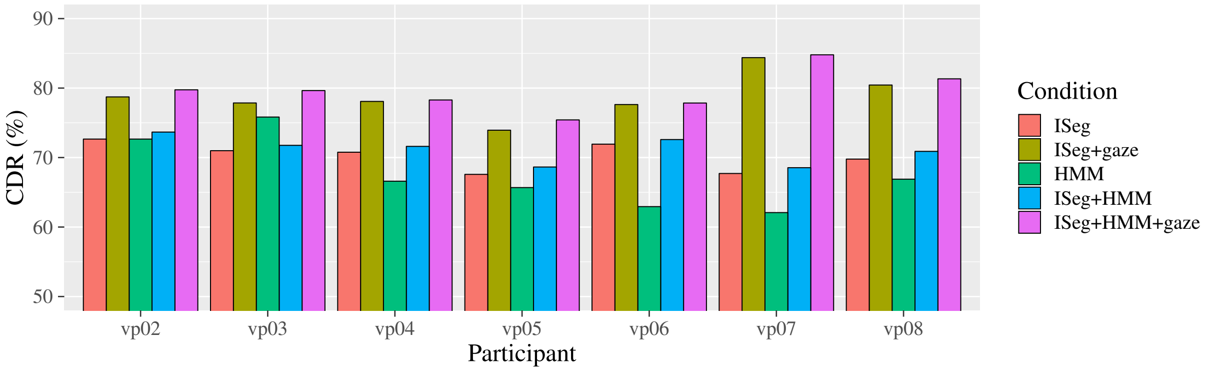

Figure 2 – The correct detection rate (CDR) for different conditions is shown for all participants. Out of all conditions, *ISeg+HMM+gaze* gives the best CDR.

of gaze data. For a 16 kHz speech signal, having a window size of 8 ms and an overlap of 4 ms, each column of the spectrogram corresponds to 4 ms of speech as specified by Leow et al. [12].

In order to collect the gaze data for the same segment of speech, the signal was divided into frames of 4 ms to ensure that the gaze data and the spectrogram column belong to the same segment of speech. This also ensures that the number of columns in the spectrogram and the number of frames of the gaze data are the same. We computed for each frame its accumulated fixation duration by summing up all fixation durations that occurred in the corresponding frame. The fixations from the *spectrogram* region were considered for this purpose. The results were assembled in a vector $\vec{F}$, such that the entry f_i represents the accumulated fixation duration for frame i. To apply the gaze data to S, the vector $\vec{F}$ is converted into a weight vector $\vec{W}$ whose entries w_i are computed as follows:

$$w_i = \begin{cases} 0.7 & \text{if} \quad f_i = 0 \\ 1.0 & \text{if} \quad 0.7 < f_i < \text{mean}(\vec{F}) \\ 1.2 & \text{if} \quad \text{mean}(\vec{F}) < f_i < 2 \times \text{mean}(\vec{F}) \\ 1.5 & \text{if} \quad f_i > 2 \times \text{mean}(\vec{F}) \end{cases} \tag{5}$$

The weights are applied to the columns of the spectrogram, producing a new spectrogram S':

$$S' = (\text{diag}(\vec{W})S^{\top})^{\top} \tag{6}$$

These weights were obtained by trying out different settings and selecting the weights that led to the best results. A small value of w_i reduces the intensity of that column of the spectrogram, and vice versa. The spectrogram S' is normalized to bring the intensity values to the range $[0, 255]$. Finally, a median filter [24, pp. 469–476] is applied, with the same settings as [12], to remove any noise. The modified spectrogram S' is processed for segmentation according to the ISeg algorithm. Table 3 summarizes the results of combining gaze data with these experimental conditions:

ISeg reference, image-only ISeg algorithm;
ISeg+gaze ISeg combined with gaze data;
HMM baseline, acoustic-only HMM segmentation;
ISeg+HMM ISeg combined with HMM segmentation;
ISeg+HMM+gaze ISeg combined with gaze data and HMM segmentation.

4.3 Fusion results

Fig. 2 shows the CDR for all conditions and for all the participants. From the figure, it is evident that the *ISeg+HMM+gaze* condition outperforms all other conditions with respect to

CDR. However, CDR alone is not a useful metric to evaluate segmentation as it does not take into account the over-segmentation that was caused by unwanted boundaries. Therefore, the results are discussed in detail with the effect of OS and the R-value.

Table 3 lists the results for all scores, over all participants and conditions. It should be noted that our corpus is different from the TIMIT corpus in several ways. First, the duration of the audio file is longer than the average TIMIT sentence; second, we noticed an audible background noise in our corpus which is not present in the TIMIT corpus. Therefore, a difference in the results is inevitable. Considering these differences, the segmentation of our data gives different results than that of the TIMIT corpus (see Table 1) with a CDR of 70% and an R-value of 0.74.

As we can see in Table 3, by comparing the performance of *ISeg* and *ISeg+gaze*, the gaze data with the ISeg algorithm improves the CDR by 8%. The OS increases, which shows that not all boundaries added by ISeg were correct. This effect is also visible in the R-value, which is reduced from 0.74 to 0.60. These results show that adding gaze data improves the performance of the ISeg algorithm, albeit at the cost of an increased OS rate.

For the HMM results, we get a low CDR. The reason for this could be the unusual length of the utterance, i.e., 46 s, as WebMAUS is generally designed for shorter utterances. The HMM segmentation has a negative mean OS of -9.37%, which is due to the fact that the manual segmentations produced by the phoneticians contain more boundaries than the HMM segmentations. When the ISeg results are combined with the HMM segmentation, an improvement is achieved in the CDR with a slightly higher OS rate. However, the R-value is not affected that much in this case.

Finally, when ISeg, HMM, and gaze are combined, the highest results for CDR (79.58%) of all conditions are obtained. This can be seen in Table 3, for the condition *ISeg+HMM+gaze*. Again, this produces a higher OS rate. The F-score for this condition stays the same (0.67%); however, the R-value is decreased because of the higher OS rate.

The results of adding gaze data to *ISeg* or *ISeg+HMM* show that weights generated from the gaze data do indeed provide a significant performance boost to the ISeg algorithm. The F-score and R-value in Table 3 show that some unwanted boundaries are also generated. One reason for this behavior is the internal working of the ISeg algorithm, as it is very sensitive to changes in intensity level [12]. By multiplying the spectrogram with the gaze data weights, the intensity values of the spectrogram change, affecting the segmentation produced by the ISeg algorithm. Also, only the fixations of the *spectrogram* region were considered for generating weights, but in practice the boundaries can also be placed by fixating on the other regions of the Praat GUI.

5 Conclusion and future work

In this paper, we have presented a gaze based modality which can be combined with existing segmentation techniques to improve the quality of automatic segmentation. In order to imitate the way human experts visually process the spectrogram representation of a speech signal during segmentation tasks, we used and analyzed suitable gaze data obtained with an eye tracker.

Furthermore, we converted the gaze data into weights and used them in an image based segmentation approach. The weights are based on the amount of time the phoneticians spent on a specific speech segment. To evaluate the usefulness of this new modality, we combined it with ISeg and HMM based segmentation. From the results, it is clear that combining the gaze data with ISeg improves the results over the baseline.

We plan to extend this research, with the goal of improving automatic speech segmentation by systematically identifying signal regions that are problematic to reliably segment in a conventional, acoustics-only approach. There are also potential applications in gaze based assistance during manual segmentation tasks. To use the system for such a task, we will need to obtain a gaze “profile” of the phoneticians during segmentation. This could be done by training a neural network model on the multimodal corpus and generalizing the gaze behavior of phoneticians, although this might also require us to record more training data. Without such a trained model, it would be difficult to apply the approach to new speech data.

6 Acknowledgments

We are grateful to Leow et al. [12] for providing us with the ISeg source code and valuable insight into their work. We would also like to thank an anonymous reviewer for valuable feedback. This study was funded by the German Research Foundation (DFG) under grant number EXC 284.

References

[1] Svendsen, T. and F. K. Soong: *On the automatic segmentation of speech signals*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 77–80. Dallas, TX, USA, 1987. doi:10.1109/ICASSP.1987.1169628.

[2] Logan, B.: *Mel frequency cepstral coefficients for music modeling*. In *International Symposium on Music Information Retrieval (ISMIR)*. Plymouth, MA, USA, 2000. URL http://ismir2000.ismir.net/papers/logan_paper.pdf.

[3] Hermansky, H.: *Perceptual linear predictive (PLP) analysis of speech*. *Journal of the Acoustical Society of America*, 87(4), pp. 1738–1752, 1990. doi:10.1121/1.399423.

[4] Khan, A., I. Steiner, Y. Sugano, A. Bulling, and R. Macdonald: *A multimodal corpus of expert gaze and behavior during phonetic segmentation tasks*. In *Language Resources and Evaluation Conference (LREC)*, pp. 4277–4281. Miyazaki, Japan, 2018. URL http://www.lrec-conf.org/proceedings/lrec2018/pdf/676.pdf.

[5] Duchowski, A. T.: *A breadth-first survey of eye-tracking applications*. *Behavior Research Methods, Instruments, & Computers*, 34(4), pp. 455–470, 2002. doi:10.3758/BF03195475.

[6] Rayner, K.: *Eye movements in reading and information processing: 20 years of research*. *Psychological Bulletin*, 124(3), pp. 372–422, 1998. doi:10.1037/0033-2909.124.3.372.

[7] Uttal, W. R. and P. Smith: *Recognition of alphabetic characters during voluntary eye movements*. *Perception & Psychophysics*, 3(4), pp. 257–264, 1968. doi:10.3758/bf03212741.

[8] Loftus, G. R. and N. H. Mackworth: *Cognitive determinants of fixation location during picture viewing*. *Journal of Experimental Psychology: Human Perception and Performance*, 4(4), pp. 565–572, 1978. doi:10.1037/0096-1523.4.4.565.

[9] Kundel, H. L., C. F. Nodine, and D. Carmody: *Visual scanning, pattern recognition and decision-making in pulmonary nodule detection*. *Investigative Radiology*, 13(3), pp. 175–181, 1978. doi:10.1097/00004424-197805000-00001.

[10] KLATT, D. H. and K. N. STEVENS: *On the automatic recognition of continuous speech: Implications from a spectrogram-reading experiment. IEEE Transactions on Audio and Electroacoustics*, 21(3), pp. 210–217, 1973. doi:10.1109/TAU.1973.1162453.

[11] DENNIS, J., H. D. TRAN, and H. LI: *Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Processing Letters*, 18(2), pp. 130–133, 2011. doi:10.1109/LSP.2010.2100380.

[12] LEOW, S. J., E. S. CHNG, and C.-H. LEE: *Language-resource independent speech segmentation using cues from a spectrogram image.* In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5813–5817. Brisbane, Australia, 2015. doi:10.1109/ICASSP.2015.7179086.

[13] GAROFOLO, J. S., L. F. LAMEL, W. M. FISHER, J. G. FISCUS, D. S. PALLETT, N. L. DAHLGREN, and V. ZUE: *TIMIT Acoustic-Phonetic Continuous Speech Corpus.* 1993.

[14] AVERSANO, G., A. ESPOSITO, and M. MARINARO: *A new text-independent method for phoneme segmentation.* In *IEEE Midwest Symposium on Circuits and Systems (MWSCAS)*, vol. 2, pp. 516–519. 2001. doi:10.1109/mwscas.2001.986241.

[15] ESPOSITO, A. and G. AVERSANO: *Text independent methods for speech segmentation.* In G. CHOLLET, A. ESPOSITO, M. FAUNDEZ-ZANUY, and M. MARINARO (eds.), *Nonlinear Speech Modeling and Applications*, pp. 261–290. Springer, 2005. doi:10.1007/11520153_12.

[16] ESTEVAN, Y. P., V. WAN, and O. SCHARENBORG: *Finding maximum margin segments in speech.* In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 937–940. Honolulu, HI, USA, 2007. doi:10.1109/ICASSP.2007.367225.

[17] BOERSMA, P.: *Praat, a system for doing phonetics by computer. Glot International*, 5(9/10), pp. 341–345, 2001.

[18] INTERNATIONAL PHONETIC ASSOCIATION: *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* Cambridge University Press, 1999.

[19] KVALE, K.: *Segmentation and Labelling of Speech.* Ph.D. thesis, Norwegian University of Science and Technology (NTNU), 1993. URL http://hdl.handle.net/11250/2368838.

[20] FLEISS, J. L.: *Measuring nominal scale agreement among many raters. Psychological Bulletin*, 76(5), pp. 378–382, 1971. doi:10.1037/h0031619.

[21] LANDIS, J. R. and G. G. KOCH: *The measurement of observer agreement for categorical data. Biometrics*, 33(1), pp. 159–174, 1977. doi:10.2307/2529310.

[22] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services. Computer Speech & Language*, 45, pp. 326–347, 2017. doi:10.1016/j.csl.2017.01.005.

[23] RÄSÄNEN, O., U. LAINE, and T. ALTOSAAR: *An improved speech segmentation quality measure: the R-value.* In *Interspeech*, pp. 1851–1854. Brighton, UK, 2009. URL https://www.isca-speech.org/archive/interspeech_2009/i09_1851.html.

[24] LIM, J. S.: *Two-dimensional Signal and Image Processing.* Prentice-Hall, 1990.