THE INFLUENCE OF SIGNAL SEGMENTATION METHODS ON RHYTHM-BASED SPEAKER RECOGNITION

Neda Mousavi[1], *Sven Grawunder*[1,2]

[1]*Martin Luther University Halle-Wittenberg, Germany,* [2]*Max Planck Institute for evolutionary Anthropology, Leipzig, Germany*
neda.mousavi@sprechwiss.uni-halle.de, sven.grawunder@sprechwiss.uni-halle.de

Abstract: This study investigates the effects of speech segmentation methods on speaker recognition models, particularly with regard to the use of rhythmic feature sets. Using three automatic methods and one manual method on the German database of Kiel corpus, segmentation was performed based on the identification of vowel onsets. Subsequently rhythmic variability indices derived from these intervals were calculated and used for principal component analysis and support vector machine model in order to investigate the variation among speakers. The results underline the influence of signal segmentation methods on speaker recognition models.

1 Introduction

This study focuses on the central role of signal segmentation methods in the extraction of rhythmic features and their impact on the accuracy of speaker recognition models. Our main focus is on the methodological subtleties of data preparation. While recognizing the importance of modeling speaker variability, this research specifically explores the nuances of raw signal segmentation and rhythmic feature extraction for such models. The findings have the potential to improve our understanding of how to refine data preparation techniques in rhythm-based speaker recognition models, and ultimately increase the reliability and effectiveness of these models.

The enhancement of speech segmentation methods remains an important area of ongoing research. Manual methods can lead to limitations due to their resource-intensive and subjective nature. The process could take up to 400 times the real-time duration, or an average of 30 seconds per phone [13]. Therefore, the focus of research has shifted to automatic methods. These include various strategies, including acoustic-driven segmentation, which aims to divide large audio files into smaller sections, such as sentences or pause intervals, especially for applications such as automatic translation [5]. With regard to automatic phonemic segmentation, two main aspects are mentioned in the literature. Forced alignment methods that are based on orthographic transcription and require transcription assistance in order to annotate audio files based on the spoken content. However, there are other text-independent methods based on self-supervised ASR models that do not rely on providing textual content [28]. Moreover, specific psycholinguistic approaches incorporate cues from the internal metrical structure of words and focus on identifying perceptually relevant phonetic events such as word boundaries, positions of heavy syllables or vowel onsets for segmentation purposes. These approaches can provide a theoretical basis for the segmentation of speech based on the metrical structure of speech, independent of phonemic content [18].

This study was motivated by our previous research [17] in which we investigated the accuracy of vowel onset identification using four segmentation methods - one manual and three

DOI: 10.35096/othr/pub-7101, URN: urn:nbn:de:bvb:898-opus4-71014

automatic methods, including WebMaus service [11], automatic segmentation based on amplitude envelope extraction and automatic segmentation based on amplitude envelope derivation [14]. The first automatic segmentation method relies on orthographic transcription of audio files to extract segmentations at the word and phoneme levels. The other two methods use acoustic cues from the amplitude envelope to detect vowel onsets as part of the metrical structure of speech. Based on the time of the occurrence, the segmentation could be performed as vowel onset intervals. However, this study takes a different direction from our earlier focus by looking at the subtle differences between rhythmic indicators identified by the determined intervals. In particular, it examines their performance in detecting between-speaker variation based on identified vowel onsets.

2 Data

Our study uses a restricted dataset consisting of German read speech extracted from the Kiel corpus [12]. The dataset comprises eight speakers, evenly distributed between males and females. To ensure enough amount of data per speaker, we strategically divided the read text into three different sections: Beginning, Middle, and End. This division was made with the primary aim of augmenting the data set to improve the accuracy and reliability of the subsequent analyses while mitigating the challenge of class imbalance.

3 Segmentation Methods

The raw speech data was segmented by a combination of manual segmentation and the use of three different automatic algorithms. The units of segmentation primarily target the vowel onsets. We evaluated four segmentation approaches: manual segmentation (Manual) as a reference, automatic segmentation by WebMaus, segmentation based on amplitude envelope extraction (AM), and segmentation derived from amplitude envelope derivation (AM-der). The manual segmentation was achieved by revising the *Praat* [2] TextGrids of the Kiel corpus and applying manual corrections. The automatic segmentation using the WebMaus system [11] extracted another version of Praat TextGrids. A Praat script then processes the extracted TextGrids specifically to determine the vowel onset intervals. The next two automatic methods were based on the extraction of the amplitude envelope. In the following section, we first give a comprehensive overview of segmentation methods based on amplitude envelope extraction and then present our approach for this study.

3.1 Speech segmentation based on amplitude envelope extraction

The study of temporal patterns of speech using features derived from the amplitude envelope (also called the speech envelope, temporal envelope, intensity contour, and intensity profile over time) has its roots in long-standing "modulation theory" [20]. According to this theory, speech sounds are produced by carrier signals generated by the vocal folds. The amplitude and frequency of these signals are subject to dynamic changes caused by shifts in the vocal tract during phonation [27]. While certain methods for extracting envelope information from speech signals have their roots in the 1980s and 1990s (e.g. Schotola [25]), the use of this approach to study speech rhythm and temporal patterns has only gained significant recognition in the last two decades.

The amplitude envelope represents a smoothed signal that captures gradual amplitude modulations within the speech waveform and maps temporal envelope fluctuations over time. Numerous studies in the literature provide a comprehensive overview of different extraction meth-

ods (see [10], [14]). The "envelope follower" is a widely used method for estimating the amplitude envelope of a signal. In this technique, the signal is rectified by taking its absolute values and then applying a low-pass filter to achieve smoothing [21]. Another approach is to calculate the root mean square (RMS) of the waveform using a moving window with finite support [7]. Applying the Hilbert transform to derive the analytical signal for estimating the amplitude envelope [1] is an additional technique. The Hilbert transform is a mathematical technique for extracting the envelope of a signal by decomposing it into its analytical and complex components. Again another method is to use prominent peaks in the signal directly to interpolate and generate the envelope [15]. And eventually a more advanced method, called cepstral smoothing [3], uses the real cepstrum obtained by the inverse Fourier transform of the logarithmic magnitude spectrum. Each coefficient of the cepstrum corresponds to the energy within certain frequency bands of the magnitude spectrum of the signal. Applying a low-pass filter to the cepstrum refines the magnitude spectrum, and it is recommended to set the cut-off frequency below the 'period'[10] of the signal to display only the spectral envelope and eliminate partial information. Empirical mode decomposition (EMD) is an alternative method to envelope extraction that involves an optimization process that minimizes a quadratic cost function. This function serves as a mathematical measure of how effectively the extracted envelope model represents the original signal. EMD, developed by Huang et al. [9], is an adaptive technique designed to analyze nonlinear and nonstationary data [10].

In this study, the methodology proposed by Tilsen and Arvaniti [26] is applied. First, the speech signal is subjected to band-pass filtering using a fourth-order Butterworth filter in the frequency band [400, 4000] Hz. This step aims to attenuate the effects of the fundamental frequency (F0) and reduce artifacts stemming from high-frequency components such as sibilants and bursts that could potentially affect the envelope peaks. Subsequently, chunk extraction becomes a central procedure to prepare the signal for amplitude envelope extraction. This process involves determining the optimal length of the speech intervals, a parameter that is influenced by factors such as recording quality and speech rate. It is important to find a balance in order to avoid too short intervals that contain too little information or too long intervals that lead to unwanted variability. To maintain consistency of rhythm and ensure the integrity of the analysis, it is advisable to minimize pauses within speech intervals and carefully consider overlaps between sections. In this study, the signal is segmented into silence and speech, with the pause intervals in Praat determined using a script based on the "To TextGrid (Silences)" command. Consequently, overlaps between the intervals are not considered necessary for the analysis. The identification of the vowel onsets was performed by determining the time points of the peak positions in both the amplitude envelope and its derivative, as described in [14]. To accomplish all these tasks, a Praat script developed by [8], incorporating the Hilbert transform, was considered. The script was adapted for use in the R environment by [22]. We have further developed the code to derive the first derivative of the amplitude envelope and to identify the peak positions within the envelope and its derivative.

3.2 Feature extraction

As our signal segmentation focuses on the vowel onset intervals, we focus on identifying rhythm indicators that align with these units. Several indicators have been proposed for the time domain, including the number of identified vocalic intervals, the standard deviation of vocalic intervals [23], the normalized pairwise variability index for vowel intervals (nPVIv) [6], and the rate normalized standard deviation of vocalic intervals (Varco) indices [4]. Moreover, in research by [19], VtoV has gained attention as a relevant metric that encompasses the mean distance between successive vowel onsets. This metric can capture the temporal dynamics and

distribution of vowel components within speech, and thus provide valuable insights into the rhythmic patterns surrounding vowel onset intervals.

Rhythmic measurements related to vowel intervals go beyond the time domain and include additional indicators in the domains of intensity and frequency. In the intensity domain, as recommended by He and Dellwo [8], we examined features related to both mean and maximum intensity variations across vowel onset intervals (including stdevM, stdevP, varcoM, varcoP, nPVIm, nPVIp for the mean (M) and peak (P) of intensity in each interval). In addition, we extended our investigation to analog attributes in the frequency domain, following the approach described by [16]. In [8], the calculations of the metrics were based on syllable units. In our study, we replace this approach with vowel onset intervals, which we assume to be generally consistent. Based on this, we derived 12 different rhythmic metrics from the data set.

4 Results

4.1 PCA Analysis of Rhythmic Features across Different Segmentation Methods

In this study, we applied principal component analysis (PCA) to condense the original feature set, with the aim of highlighting key rhythmic details while addressing concerns about inter-correlation and overfitting. The main objective is to assess the separability of observations, which are based on individual speakers' extracted feature sets using different segmentation methods, and further, to investigate the influence of the segmentation methods on the overall analysis. For visualizing the clusters, circular outlines were used, each centered around the centroid of a particular segmentation type (s. Fig 1).

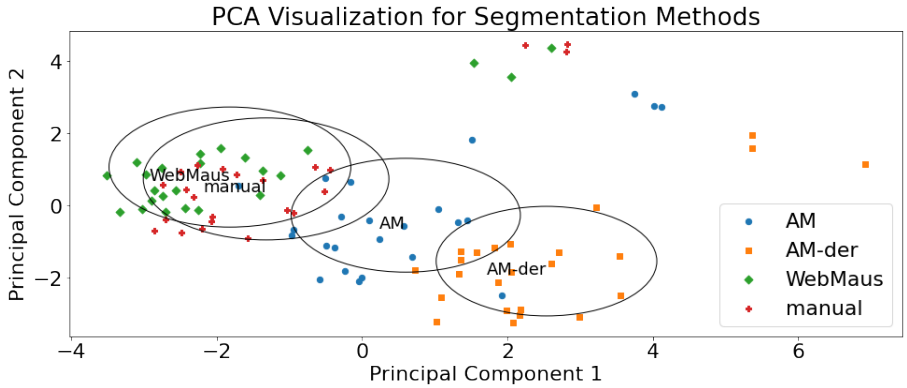

Figure 1 – PCA representation across Different Segmentation Methods

The different regions in the PCA representation show the differentiation of the observations based on different segmentation methods. The separability of these regions and the extracted clusters underlines the influence of the segmentation method on the feature space. To quantify separability, we use the Silhouette Score (SilSc) [24], a metric that objectively measures the separation and distinctness of clusters. This score takes into account both the cohesion of points within clusters and the separation between different clusters, thus providing a comprehensive measure of the degree of separation. A higher SilSc approaching 1 indicates well-defined and distinct clusters, while a lower score indicates some degree of overlap between clusters. Negative values in the SilSc may indicate possible misclassification or insufficient separation between clusters. In our study, the SilSc plays a crucial role in assessing the degree of separation or overlap between observations derived from different segmentation types and provides insights into potential similarities in the rhythmic features captured.

The SilSc determined for the individual segmentation types indicate different degrees of

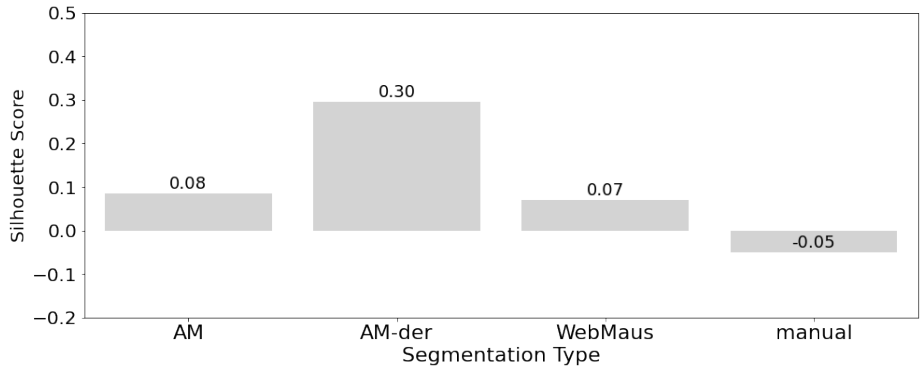

Figure 2 – Silhouette Scores (SilSc) for each segmentation type

cluster separation. In particular, the higher SilSc for the AM-der type indicate discrete feature space. Conversely, the lower SilSc for other types indicate a certain overlap of the extracted features, especially between WebMaus and manual types. This overlap implies a certain degree of similarity in the rhythmic features captured by these two segmentation methods. It is important to note that this analysis does not address the strengths or weaknesses of the methods, but emphasizes the distinctiveness of observations based on rhythmic features derived from different segmentation approaches.

After analyzing clustering tendencies through the SilSc, we shift our focus to assessing within-cluster variability, particularly with respect to the ability of each method to represent variation between individual speakers. Our attention turns to understanding the variability captured by the first principal component (PC1). This step is crucial to identify the specific characteristics of the speakers within each cluster and to gain insights into how well the clusters represent the individual variations present in the data.

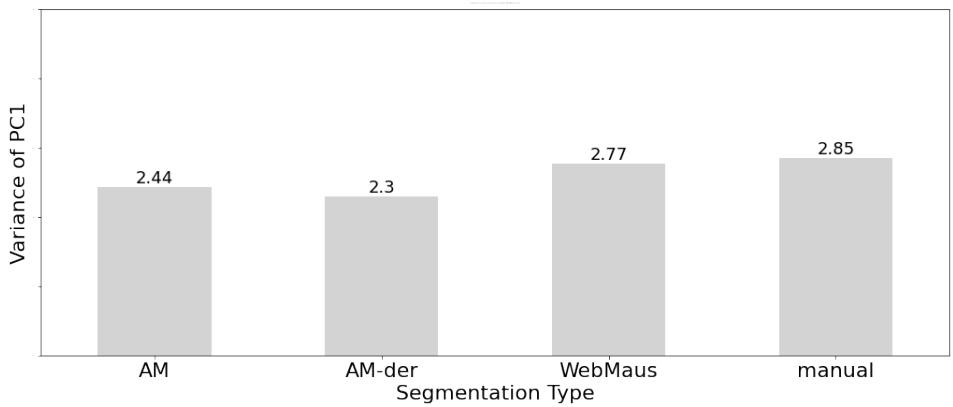

Figure 3 – The variance of the first principal component (PC1)

In examining the variances of principal component 1 (PC1) for various segmentation methods, we found notable differences. Manual segmentation had the highest variance (2.85), indicating a wider range of observations within the cluster. WebMaus method was close behind with a variance of 2.77, indicating comparable variability. In contrast, AM and AM-der had a lower variance (2.44 and 2.30 respectively), indicating a more concentrated set of points within their respective clusters. To accurately assess the significance of these differences in variance, an analysis of variance (ANOVA) was performed. The results of the ANOVA (F-statistic: 1372.49, p-value: 1.76e-15) confirmed the statistical significance of the observed variances and provided solid evidence of the heterogeneity of variances across clusters associated with different segmentation methods.

4.2 Speaker Classification using SVM model

To further investigate the effectiveness of different segmentation methods in capturing rhythmic features for speaker variation modeling, we employed the Support Vector Machine (SVM) model. Datasets derived from different segmentation methods were subjected to a comparative evaluation to measure their performance in speaker classification. Figure 4 shows the accuracy and precision achieved by each segmentation method.

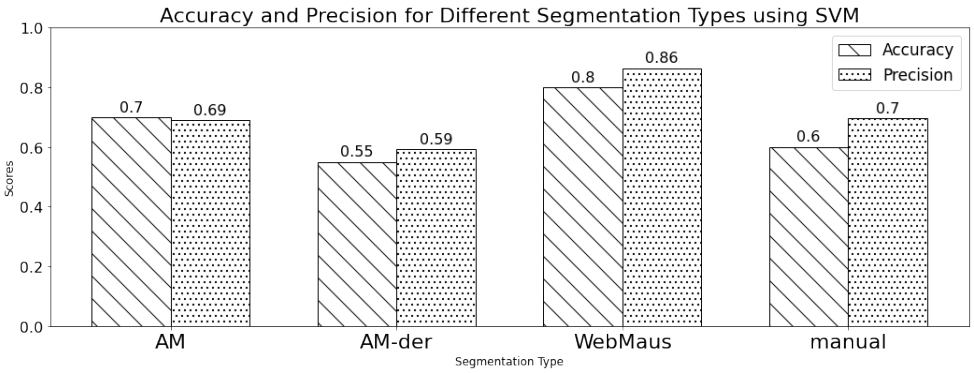

Figure 4 – Accuracy and Precision for Different Segmentation Methods using SVM Model

The segmentation methods lead to different levels of performance in capturing and characterizing the underlying rhythmic patterns for speaker classification, which was reflected in their respective accuracy and precision metrics. In particular, AM segmentation showed a remarkable accuracy of 70%, accompanied by a precision of 69%. AM-der segmentation, on the other hand, showed a comparatively lower accuracy of 55% and a precision of 59%. WebMaus segmentation showed excellent performance, with an accuracy of 80% and a high precision of 86%. Finally, manual segmentation, which requires manual intervention, achieved an accuracy of 60% and a precision of 70%.

5 Discussion

In this study we aimed for an evaluation of the effectiveness of one manual and three automatic segmentation methods in identifying rhythm-related events, particularly vowel onsets, and subsequently use these events to classify speakers based on extracted rhythmic features. Despite the widespread use of the automatic segmentation tool WebMaus, its reliance on a forced alignment approach and the need for a corresponding transcript can be challenging, especially for large datasets or languages with low resources. While there are alternative automatic methods for transcript extraction, such as ASR in the BAS WebService or self-supervised models such as Wav2Vec2, their accuracy and the potential need for manual adjustments pose significant challenges that need to be carefully considered. In this context, it is crucial to explore alternative automatic segmentation methods, especially those based solely on audio input, such as segmentation based on amplitude envelope extraction. The results suggest that the accuracy of these methods can compete with manual methods or WebMaus segmentation in identifying speakers. It is important to mention that in another study [17], we tested the accuracy of these methods in identifying vowel onset within a tolerance window of 40 ms, where the manual methods showed higher accuracy in finding vowel onsets compared to the others. However, our study here focuses specifically on the accuracy of speaker identification model based on rhythmic features derived from the vowel onsets. It is important to distinguish between these two aspects and not to mix them. It should also be noted that in this study, we focused on the analysis of German

read speech. While the results from this particular context provide valuable insights, it is important to recognize that the generalization of these results may have limitations. Therefore, to achieve broader applicability and validate the results, it is essential to investigate and include other speaking styles in future studies.

References

[1] M Florencia Assaneo, Pablo Ripollés, Joan Orpella, Wy Ming Lin, Ruth de Diego-Balaguer, and David Poeppel. Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature neuroscience*, 22(4):627–632, 2019.

[2] Paul Boersma and David Weenink. Praat: doing phonetics by computer [Computer program], 2022.

[3] Marcelo Caetano and Xavier Rodet. Improved estimation of the amplitude envelope of time-domain signals using true envelope cepstral smoothing. In *Inter Conf on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4244–4247. IEEE, 2011.

[4] Volker Dellwo, P Karnowski, and I Szigeti. Rhythm and speech rate: A variation coefficient for deltac. 2006.

[5] Ryo Fukuda, Katsuhito Sudoh, and Satoshi Nakamura. Speech segmentation optimization using segmented bilingual speech corpus for end-to-end speech translation. *arXiv preprint arXiv:2203.15479*, 2022.

[6] Esther Grabe and Ee Ling Low. Acoustic correlates of rhythm class. *Laboratory phonology*, 7(515-546), 2002.

[7] John Hajda. A new model for segmenting the envelope of musical signals: The relative salience of steady state versus attack, revisited. In *Audio Engineering Society Convention 101*. Audio Engineering Society, 1996.

[8] Lei He and Volker Dellwo. The role of syllable intensity in between-speaker rhythmic variability. *Intern J of Speech, Language & the Law*, 23(2), 2016.

[9] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. A.*, 454(1971):903–995, 1998.

[10] Cecilia Jarne. A heuristic approach to obtain signal envelope with a simple software implementation. *Anales AFA*, 29, 07 2018.

[11] Thomas Kisler, Uwe Reichel, and Florian Schiel. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347, 2017.

[12] Klaus J Kohler. Labelled data bank of spoken standard german: the kiel corpus of read/spontaneous speech. In *Proceeding of 4th Int Conf on Spoken Language Processing. ICSLP'96*, volume 3, pages 1938–1941. IEEE, 1996.

[13] Hong Leung and V Zue. A procedure for automatic alignment of phonetic transcriptions with continuous speech. In *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 73–76. IEEE, 1984.

[14] Alexis Deighton MacIntyre, Ceci Qing Cai, and Sophie K Scott. Pushing the envelope: Evaluating speech rhythm with different envelope extraction techniques. *JASA*, 151(3):2002–2026, 2022.

[15] Qinglin Meng, Meng Yuan, Zhenya Yang, and Haihong Feng. An empirical envelope estimation algorithm. In *2013 6th Intern Congr on Image and Signal Processing (CISP)*, volume 2, pages 1132–1136. IEEE, 2013.

[16] Neda Mousavi and Sven Grawunder. The classification of speaker and prosodic peculiarities in emotional speech based on rhythmic patterns. In *14th International Conference of Experimental Linguistics (ExLing 2023)*, Athens, Greece, 2023.

[17] Neda Mousavi and Sven Grawunder. Performance evaluation of speech segmentation algorithms based on amplitude envelope extraction techniques. In TR Pistor, C Steiner, F Tomascheck, and A Leemann, editors, *Book of Abstracts der 19. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, Bern, 2023. Bern Open Publishing.

[18] Dennis Norris, James M McQueen, and Anne Cutler. Competition and segmentation in spoken-word recognition. *J of Experimental Psychology: Learning, Memory, and Cognition*, 21(5):1209, 1995.

[19] Massimo Pettorino, Marta Maffia, Elisa Pellegrino, Marilisa Vitale, Anna De Meo, et al. Vtov: A perceptual cue for rhythm identification. In *Proc of the Prosody-Discourse Interface Conference 2013 (IDP-2013)*, pages 101–106. University of Leuven (KU Leuven), 2013.

[20] Reinier Plomp. *Perception of Speech as a Modulated Signal*, pages 29–40. De Gruyter Mouton, Berlin, Boston, 1984.

[21] Alexandros Potamianos and Petros Maragos. A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation. *Signal processing*, 37(1):95–120, 1994.

[22] W. Pouw and J. P. Trujillo. Extracting a smoothed amplitude envelope from audio, 2021.

[23] Franck Ramus, Marina Nespor, and Jacques Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3):265–292, 1999.

[24] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J of Computational and Applied Mathematics*, 20:53–65, 1987.

[25] Thomas Schotola. On the use of demisyllables in automatic word recognition. *Speech communication*, 3(1):63–87, 1984.

[26] Sam Tilsen and Amalia Arvaniti. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *JASA*, 134(1):628–639, 2013.

[27] Léo Varnet, Maria Clemencia Ortiz-Barajas, Ramón Guevara Erra, Judit Gervain, and Christian Lorenzi. A cross-linguistic study of speech modulation spectra. *JASA*, 142(4):1976–1989, 2017.

[28] Bryce Wohlan, Duc-Son Pham, Kit Yan Chan, and Roslyn Ward. A text-independent forced alignment method for automatic phoneme segmentation. In *Australasian Joint Conference on Artificial Intelligence*, pages 585–598. Springer, 2022.