Interspeech 2022
18-22 September 2022, Incheon, Korea

Exploring audio-based stylistic variation in podcasts

Katariina Martikainen[1,2], *Jussi Karlgren*[2], *Khiet P. Truong*[1]

[1]University of Twente, Human Media Interaction, Enschede, The Netherlands
[2]Spotify, Stockholm, Sweden

martikainen.kata@gmail.com, jkarlgren@spotify.com, k.p.truong@utwente.nl

Abstract

Podcast are a growing spoken medium that is listened to for various reasons in various situations (e.g., for entertainment or educational purposes, on the train or at home) consisting of various types of audio such as (unstructured) speech, music, and other sounds. Traditionally, search and recommendation of spoken content focuses on topical content, derived from text transcriptions, ignoring paralinguistic aspects in spoken language. Instead, we propose to model these paralinguistic aspects, such as speaking style, in podcasts to address both the heterogeneity of type of audio in podcasts and user needs to enable enriched access to this medium. In this paper, we take a first step towards this goal and explore audio-based stylistic variation in podcasts by 1) investigating what facets of stylistic variation are salient and of interest to listeners, and 2) gathering more insights into the kind of stylistic variation that is currently feasible to model with open-source audio tools and that is present in podcasts. We find that much of the stylistic variation mentioned by the users is related to speaking style and music, and we show, using open-source tools, how audio-based stylistic aspects vary across episodes, shows, and genres.

Index Terms: paralinguistics, speaking style, podcasts, spoken document retrieval

1. Introduction

The amount of new types of digitally stored casual speech in the form of interviews, lectures, debates, radio talk show archives, and educational podcasts is increasingly available [1]. One of the forms of audio media which has really gained popularity during the past years are podcasts [2]. As of 2019, there are 90 million monthly podcast listeners in the United Sates, twice as many as in 2015. Podcast usage is driven by a younger generation looking for information, entertainment, and distraction [3]. Podcasts are a particularly appealing form of entertainment when the listener's visual attention is required elsewhere. They are also an attractive option for pastime when for example commuting, cleaning, or exercising [4]. When it comes to format, podcasts are a more free form of spoken audio than for example traditionally studied broadcast news. Podcasts can come in the form of interviews, informal chats, debates, stories, and much more. In addition to this, podcasts can contain many different kinds of sound effects and music. This rich format poses a challenge as to how to make this richness accessible to users.

Access to the spoken content is traditionally enabled by automatic speech recognition (ASR) technology that provides text transcripts allowing for topical search and recommendation. However, the transcripts provided by ASR typically do not capture the richness of stylistic information carried by audio-media (ASR focuses on recognizing what is being said, not how it is being said). Users might want to, for example, search for stylistically similar content, or would benefit from recommendations, which in addition to topic consider the stylistic characteristics of the recommended media. Therefore, there is a need to address "*how*" spoken content documents are in addition to "*what*" is said in them.

While the popularity of podcasts is increasingly growing, podcasts suffer from the same challenges as other spoken content. Spoken content does not afford the same sort of skimming through or jumping back and forth as text does. Additionally, the browsing is limited in terms of speed, as a recording is more limited by the recorded speed of talking. When it comes to enabling users to find relevant content to consume and to building scalable spoken content retrieval systems, this is still a challenge both 1) from a user ends' perspective: **What are relevant spoken stylistic dimensions other than topic-related aspects that users are interested in**, and 2) from a technical perspective: **How do these stylistic dimensions vary across shows and categories/genres of podcasts**? In order to address these challenges, we first carried out user workshops, and subsequently used those results to analyze audio-based stylistic features in podcasts using Principal Component Analysis (PCA) and k-means clustering.

2. Related work

2.1. Audio and spoken content analysis

As podcasts consist of different kinds of audio, analysis of podcasts can be considered a specific form of audio content analysis. Audio content analysis typically consists of segmenting and classifying the audiostream into different types of audio events and/or speakers [5, 6]. Main audio events include speech, music, silence, environmental sounds or nonspeech [5]. Depending on the type of audio collection, detectors have also been developed for retrieval of highlights such as sport highlights [7], or vocalized behavior such as "hot spots" and laughter in meetings [8, 9]. Other social interactional aspects such as emotion, stance (e.g., disagreement, positivity [10, 11, 12], speaking styles [13]) are heavily studied as well and can be relevant key descriptors in audio collections containing dialogue, or more generally, speech. This is particularly relevant for podcasts which are typically primarily spoken-word media, mixed to a lesser extent with music, and sound effects, but it remains a question what kind of audio descriptors are relevant for podcasts, most notably for the purposes of making them findable to their intended audience [14].

2.2. Podcasts

Podcasts are a relatively new medium that have some unique characteristics different from more traditional media such as music or text documents. They are often characterised by heterogeneous spoken audio which can vary widely in quality and type; they can contain informal (multiparty) chats, monologues,

10.21437/Interspeech.2022-10871

Descr. category	Scale	Examples of sticky notes the participants grouped under categories
Music and sounds	pleasant – unpleasant	background music creates an atmosphere, music is chaotic and all over the place, using repetitive background music too long, no sound effects/music
Ads/commercials	suitable – unsuitable	ads present, commercials at the start of the podcast are irritating
Audio quality	good – bad	echo, noise in recording, well recorded and good mix, well "audio quality"
Speaking style and voice qualities	pleasant – unpleasant	nice, deep and clear voice of speaker, whiny voice, excitement of the presenter, speaker talks in a lazy/sleepy way
Formality level	serious – leisure	light mood, dynamic between 2 people (seem like friends, relaxed atmosphere), serious
Engagement level	engaging – not engaging	excitement, showing emotion, happy voices → positive mood/engaged, energy makes it possible to feel like being there in the conversation → entertained, speech too slow → annoyed
Format	suitable – unsuitable	it just starts abruptly, different personalities/attitudes, topic not clear, good and sweet end

Table 1: *Descriptive categories extracted from the workshops (exericses 2 and 3).*

interviews, lectures, and meditation in addition to background noises and music which brings along not only large topical variation but also speaking style variation. And they are listened to for many different reasons by users, e.g., for entertainment, educational or informational purposes. While some of the search tasks in podcasts might be similar to traditional information search based on topic, the goals and search strategies for podcast search (and recommendation) may be strongly influenced by the perception that users have of the capabilities of the current audio search and recommendation systems [2]. The rather specific nature of podcasts asks for (and affords) novel enriched, and more fine-grained search and recommendation capabilities. How this can be achieved and what kind of representation enables this enriched and personalized access is one of the current research challenges in podcast research.

While stylistic varation in other content such as text materials on the internet [15, 16, 17] has been investigated, and while speaking style in the context of paralinguistic and prosodic research has been addressed before [13], spoken stylistic variation in podcasts has much less often been a focus in previous research. In the PodCred framework [18, 19], the speech and style of the podcaster are deemed important factors that influence listener perceptions of the credibility and quality of podcasts. More specifically, paralinguistic features such as fluency, speech rate, use of conversational style, presence of affect were mentioned as elements that make it possible to capture the potential appeal of the podcaster's persona and also the basic ease-of-listening of the podcast. Other related work includes that of Yang et al.'s (2019) [4] who investigated features or representations that are predictive of non-textual podcast characteristics in podcasts, i.e., seriousness and energy. In short, given the few related studies we found, there is much room for further investigation into stylistic variation in podcasts, both from a user's and computational perspective.

3. Users' perception of stylistic variation in podcasts

In order to find out what audio-based stylistic dimensions are salient and interesting to listeners, we carried out three user workshops.

3.1. Methodology

Participants In total, nine participants (2m,7f) in the range of 20–29 years old from various cultures and countries joined the workshops. The frequency of podcast listening among participants varied from daily listening to once a month.

Data Fourteen example podcast episodes[1] were selected from Spotify's streaming platform for the workshops. This set of episodes was selected by the authors to ensure a heterogeneity in style, topic, and genre.

Method Due to the COVID-19 pandemic, the workshops were conducted online using video conferencing and Mural (https://www.mural.co/), an online visual collaboration tool simulating brainstorming with sticky notes. Each workshop consisted of 3 exercises, lasted between 2–3 hours and was conducted with 3 participants. In the first exercise, participants were asked to individually listen to each podcast episode and to make note of things they liked (pink notes) and disliked (blue notes) (20min in total) that were **not** related to the topic or content of the episode. After the individual listening sessions, the participants shared their observations with each other and the first author, and could clarify the sticky notes if they wanted. The first author gathered all the sticky notes and performed a thematic analysis, grouping the sticky notes together. In the second exercise, participants were asked to group together the sticky notes from exercise one into (higher-level) categories if possible and assign names to those categories. In the third exercise, we added sticky notes that the authors came up with based on literature and asked the participants to add these notes to the existing categories if possible. New categories could be formed as well. The first author manually checked the categories generated by the participants and combined them when categories looked similar and were overlapping.

3.2. Results

The results of the thematic analysis carried out for exercise 1 can be seen in Fig. 1. The participants appear to be mostly attentive to stylistic aspects related to talking: a closer inspection of the sticky notes revealed that speech rate, tone of voice, monotonousness, and articulation were mentioned in that theme. Other themes that were noted include aspects related to music, audio quality or ads. We can also observe that participants have individual preferences, although the presence of ads

[1]Economist Radio, Stuff you should know, 99%Invisible, The Tim Ferriss Show, Great Women of Business, VIEWS with David Dobrik and Jason Nash, Anna Faris Is Unqualified, The Daily Show With Trevor Noah, Everything is Alive, Sleep with me, Knifepoint horror

is disliked by almost all.

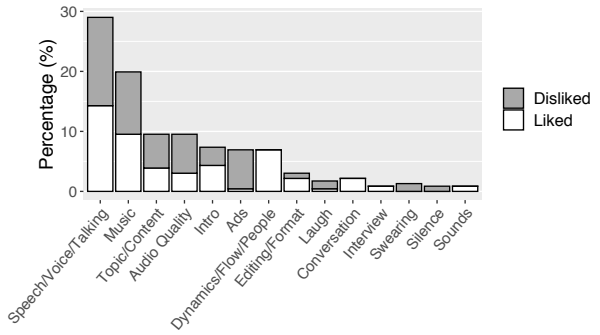

Figure 1: *Stylistic features extracted from exercise 1 (based on 231 observations written on sticky notes).*

The results of exercise 2 and 3 are presented in Table 1. The participants were able to group the sticky notes and assign names to these categories. Of particular interests are the categories "Speaking style and voice qualities", "Formality level", and "Engagement level" which are related to speaking style. In the next section, we will explore some of the acoustic features in podcasts that are related to speaking style.

4. Exploring stylistic variation in podcasts

In the following section, we investigate audio-based stylistic variation in podcasts and gauge the feasibility of characterising podcasts by their audio-based stylistic content using PCA and k-means clustering.

4.1. Methodology

Data The Spotify English Language Podcast Dataset [20] of about 100k podcast episodes was used in this study. The dataset consists of mainly English-language podcast episodes and their metadata, ranging over a wide variety of topics and includes both professional studio production and amateur material. In addition, the content is delivered in a variety of structural and discourse formats, number of speakers, and levels of formality. The data for this study was sampled from the 100k episode dataset by selecting episodes with a duration between 7 and 67 minutes and where the episode came from a show with at least 30 episodes in the dataset. This resulted in a data set of 911 episodes, see Table 2 for an overview. Very short and very long episodes were omitted since they typically were specifically oriented towards some special listening case (background ambient audio for insomniacs or pets; timing for toothbrushing; trailer advertisments for other episodes).

Method We looked for open-source tools that enabled us to extract acoustic features related to what listeners had found salient and interesting in the workshops. The inaSpeechSegmenter tool [21] seemed to fit our goals as it segments audio files into music, noise, silence, female and male speech. Note that from here on, we will refer to "female-sounding" and "male-sounding" speech as high-pitched and low-pitched voices respectively as we do not want to make any assumptions about gender classification made by a classifier not trained by the authors of this paper. The percentage of each type of segments was calculated over the whole podcast episode. Following the results of the user workshops, we also extracted mean pitch, standard deviation of pitch, and speaking rate over the speech segments that reflect speaking style to a certain extent. Although high-level categories such as formality and engagement were also mentioned in the user workshops, we leave this for future research as these require separate training to be detected. Pitch was extracted using "pYAAPT" [22] and speech rate using a Praat script by De Jong & Wempe (2009) [23]. Subsequently, PCA and k-means clustering were performed to obtain a more compact representation of stylistic variation, and to see how similarity and variation between and within shows and categories of podcasts can be characterized.

Category	**Show**
Science	Alpha Male Strategies
Society & Culture	Big Little Life with The Dashleys, Coach Corey Wayne, Heavyweight
Lifestyle & Health	Bore You To Sleep, Optimal Living Daily, Sleep and Relax ASMR, The Receipts Podcast
News & Politics	Crimetown
Education	English Speeches, Motivation and Inspiration for Ambitious Achiever
Stories	Famous Fates, Today in True Crime, Parcast Presents: March Mysteries
Music	Gynning & Berg
Religion & Spirituality	Purely Being Guided Meditations
Arts & Entertainment	The Hottest Take

Table 2: *Spotify's categorisation of the selected podcast shows.*

4.2. Results

In order to see what kind of acoustic variables could capture audio-based stylistic variation in podcasts, we carried out a principal component analysis. Based on the elbow method, we report on the first 5 PCs. Table 3 shows the correlation loadings and explained variance of the PCA obtained: with the first two PCs, 60.9% of the variance is explained. It seems that PC1 is mostly concerned with high-pitched voices and large pitch variability. PC2 seems to focus on low-pitched voiced and the amount of speech, in fact, most of the speech is low-pitched in this dataset. Music and speech rate have high loadings for PC3 and PC4, while silence is high on PC5. The first 2 PCs can also be inspected in this biplot, see Fig. 2, where the individual observations, coloured by show, are plotted as well. We can observe that shows seem to cluster reasonably well (visually), although there is some variation within each show (same for category, figure not shown here due to space limitation).

In order to explore to what extent audio-based stylistic content would yield well-formed clusters, we carried out k-means clustering on the 5 PCs. Based on the Within-Cluster-Sum of Squared Errors (WSS) and a scree plot, the number of clusters was determined to be 5. The obtained clusters are visualized in Fig. 3 in which observations are labelled by category through point type. The color indicates membership of the clusters as a result of the k-means clustering. We can observe that some clusters show little variance among category, for example, clusters 3 and 4 consists of episodes mainly from "Lifestyle & Health" and "Religion & Spirituality". But there are also clusters where categories are highly mixed, indicating that stylistic variation is present across categories (and shows). In order to assess the obtained clustering based on stylistic variation and compare it to

	PC1	PC2	PC3	PC4	PC5
music	-0.043	-0.425	**0.719**	**0.476**	-0.045
noise	-0.377	**-0.787**	-0.420	-0.029	0.039
silence	-0.530	0.429	0.319	-0.222	**0.615**
speech	0.517	**0.806**	-0.099	-0.162	-0.192
female	**0.857**	-0.180	-0.048	-0.415	0.099
male	-0.621	**0.643**	-0.004	0.351	-0.213
pitch_average	**0.838**	0.089	0.117	0.303	0.171
pitch_std	**0.901**	-0.059	0.172	0.191	0.020
speech_rate	0.196	0.142	**-0.618**	**0.626**	0.301
% variance	37.394	23.464	13.698	12.423	6.598
cum. % variance	37.394	60.858	74.556	86.979	93.577

Table 3: *Correlation loadings of the PCA components and variance explained.*

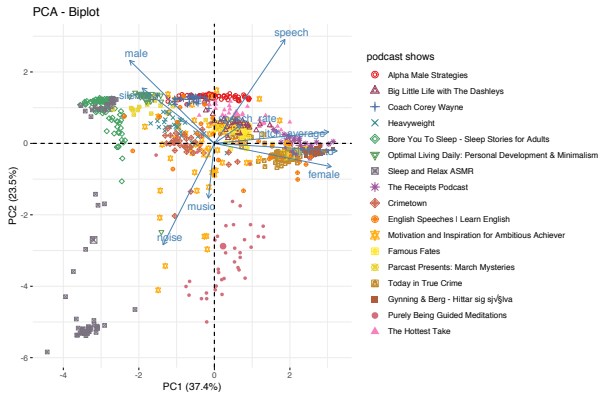

Figure 2: *Biplot with individual observations and loading vectors of PCA*

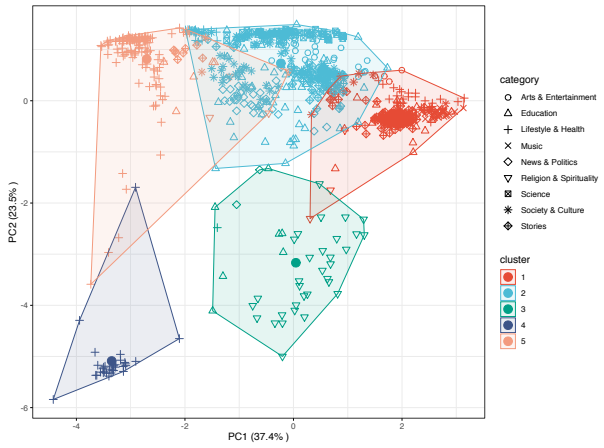

Figure 3: *Results of k-means clustering (k=5) after PCA, observations labelled by category.*

the "clustering" based on show and category, we calculated the BSS (Between Sum of Squares)/TSS (Total Sum of Squares) ratios for these 3 different types of partitioning; the higher this ratio, the higher the quality of a certain partitioning. For the calculation of BSS, we used the centroids of the clusters for the stylistic variation partition. For show and category, the means of the PCA coordinates of the different shows and categories served as centroids. The BSS/TSS ratio is highest for the stylistic variation-based clustering, followed by show and category, see Table 4.

partitioning	BSS/TSS ratio
stylistic variation	$5699.970/7672.369 = 0.743$
show	$5313.056/7672.369 = 0.692$
category	$3201.458/7672.369 = 0.417$

Table 4: *BSS/TSS ratios for different grouping types.*

5. Discussion and Conclusion

Our results contribute to the existing body of work on non-textual descriptors of podcasts [4, 19] and serves as a first step to actually operationalizing and making audio-based stylistic variation in podcasts explicit. The clustering results indicate that it is worthwhile to group podcast episodes by their audio-based stylistic content, and that it could potentially aid in personalized user search and recommendation, in addition to category and show classifications. While we have learned from the user workshops that listeners are interested and attentive of speaking styles in podcasts, it remains to be seen in future research how listeners actually make use of this information once speaking style information is accessible in search and recommendation systems. We also recommend to look on a more detailed level into podcast episodes as the interaction and speaking style can change within the episode and could allow for more fine-grained search and recommendation [14]. Finally, a major challenge lies in modelling these higher-level categories of speaking style that listeners referred to in the workshops, e.g. formality level, engagement. Pitch and speaking rate are underlying low-level acoustic features that could cue these categories but they do not grasp the full concept of formality or engagement.

In conclusion, the user workshops showed that listeners are perceptive of and interested in non-textual spoken aspects in podcasts, and are able to verbalise these as speaking style, voice qualities, engagement and formality. Using open-source tools, we explored several of these spoken stylistic features such as amount of music, noise, pitch average, pitch standard deviation, and speaking rate. To capture the stylistic variation, a PCA analysis was carried out and revealed that the first component seems to be positively correlated with pitch-related characteristics, the second component seems to be correlated with "male-sounding" speech and negatively correlated with amount of noise in the podcast. Subsequently, a k-means clustering carried out on the 5 PCs showed that this clustering based on audio-based stylistic variation has a higher BSS/TSS ratio than a partitioning based on show or genre. This suggests that audio-based stylistic variation expressed in acoustic features such as amount of music, pitch average and standard deviation, and speaking rate, is a promising representation of similarly-sounding podcasts in terms of stylistic content. Future studies should look into how listeners can benefit from access to audio-based stylistic variation in podcasts.

6. Acknowledgements

The work presented in this paper is based on the first author's master thesis "Audio-based Stylistic Characteristics of Podcasts for Search and Recommendation: A User and Computational Analysis" [24]. We wish to extend a heartfelt thank you to the participants who took part in our workshops.

7. References

[1] M. A. Hearst, *Search User Interfaces*. Cambridge University Press, 2009.

[2] J. Besser, K. Hofmann, and M. Larson, "An exploratory study of user goals and strategies in podcast search." in *Proceedings of Workshop Information Retrieval*, 2008, pp. 27–34.

[3] N. Newman and N. Gallo, "News podcasts and the opportunities for publishers," in *Digital News Report*. Oxford: Reuters Institute of Journalism, 2019.

[4] L. Yang, Y. Wang, D. Dunne, M. Sobolev, M. Naaman, and D. Estrin, "More than just words: Modeling non-textual characteristics of podcasts," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 276–284.

[5] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[6] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," in *Proceedings of the fourth ACM International conference on Multimedia*, 1997, pp. 21–30.

[7] H. Boril, A. Sangwan, T. Hasan, and J. H. Hansen, "Automatic excitement-level detection for sports highlights generation." in *Proceedings of Interspeech*, 2010, pp. 2202–2205.

[8] B. Wrede and E. Shriberg, "Spotting "hot spots" in meetings: human judgments and prosodic cues." in *Proceedings of Interspeech*, 2003, pp. 2805–2808.

[9] K. P. Truong and D. A. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.

[10] C. Lai, B. Alex, J. D. Moore, L. Tian, T. Hori, and G. Francesca, "Detecting topic-oriented speaker stance in conversational speech." in *Proceedings of Interspeech*, 2019, pp. 46–50.

[11] G.-A. Levow, V. Freeman, A. Hrynkevich, M. Ostendorf, R. Wright, J. Chan, Y. Luan, and T. Tran, "Recognition of stance strength and polarity in spontaneous speech," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 236–241.

[12] N. G. Ward, J. C. Carlson, and O. Fuentes, "Inferring stance in news broadcasts from prosodic-feature configurations," *Computer Speech & Language*, vol. 50, pp. 85–104, 2018.

[13] N. Ward, "Individual interaction styles: Evidence from a spoken chat corpus," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2021, pp. 27–31.

[14] R. Jones, H. Zamani, M. Schedl, C.-W. Chen, S. Reddy, A. Clifton, J. Karlgren, H. Hashemi, A. Pappu, Z. Nazari *et al.*, "Current challenges and future directions in podcast information access," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1554–1565.

[15] J. Dewe, J. Karlgren, and I. Bretan, "Assembling a balanced corpus from the internet," in *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*, 1998, pp. 100–108.

[16] J. Karlgren, "Textual stylistic variation: Choices, genres and individuals," in *The Structure of Style*. Springer, 2010, pp. 113–125.

[17] ——, "Conventions and mutual expectations," in *Genres on the Web*, ser. Text, Speech and Language Technology, A. Mehler, S. Sharoff, and M. Santini, Eds. Springer, 2010, vol. 42, pp. 33–46.

[18] M. Tsagkias, M. Larson, W. Weerkamp, and M. De Rijke, "Podcred: A framework for analyzing podcast preference," in *Proceedings of the 2nd ACM workshop on Information credibility on the web*, 2008, pp. 67–74.

[19] M. Tsagkias, M. Larson, and M. De Rijke, "Predicting podcast preference: An analysis framework and its application," *Journal of the American Society for information Science and Technology*, vol. 61, no. 2, pp. 374–391, 2010.

[20] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette *et al.*, "100,000 podcasts: A spoken english document corpus," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5903–5917.

[21] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[22] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[23] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[24] K. Martikainen, "Audio-based stylistic characteristics of podcasts for search and recommendation: a user and computational analysis," Master's thesis, University of Twente, 2020.