INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Matching Acoustic and Perceptual Measures of Phonation Assessment in Disordered Speech - A Case Study

Melanie Jouaiti[1,2], *Pippa Kirby*[1,2], *Ravi Vaidyanathan*[1,2]

[1] Imperial College London, UK
[2] UK Dementia Research Institute Care Research and Technology Centre, UK
m.jouaiti@imperial.ac.uk, p.kirby1@imperial.ac.uk, r.vaidyanathan@imperial.ac.uk

Abstract

Speech/voice disorders are common in People Living with Dementia (PLwD). Fluctuations in speech quality can serve as biomarkers of cognitive deterioration but there is a gap in automated assessment of speech collected in unstructured environs. Our organisation has deployed Alexa in the households of 14 PLwD to track self-reported mental and physical state as well as use of language.

In this work, we present a case study analysing highly variable speech over time, providing potential insights into cognitive changes. Alexa data gathered from the participant was manually annotated with speech assessment labels. Those labels are matched to openSMILE features by performing a feature importance analysis to isolate critical features that contribute to the perceptual ratings. We can assess phonation with a F1-score of 0.55, breathiness: 0.71, roughness: 0.60, asthenia: 0.65, strain: 0.74. This work is a first step towards automatic speech assessment to monitor cognitive impairment over time.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Dementia is a leading cause for disability and dependence in elderly across the world. People living with dementia (PLWD) require social support, day care facilities and supported residences with advancing illness. Cognitive decline in dementia-related neuro-degenerative conditions (e.g.Alzheimer's Disease) often manifest in speech and language disorders [1, 2] (See [3] for a review of language impairments) , which can appear years before other symptoms of cognitive deterioration [4, 3]. Mapping pre-symptomatic cognitive decline through non-invasive bio-markers holds significant promise for early detection and clinical monitoring of disease progression. Moreover, identifying relevant acoustic features indicative of decline supports automatic assessment of changes in speech quality over extended periods of time enables detecting disordered speech to trigger alert systems for support of populations at risk [5].

Speech/voice is typically assessed by a speech and language therapist, or computationally when subjects periodically repeat the same sentences. Voice disorders detection has been tackled using different approaches, including vocal hyperfunction, acoustic analysis approaches employing neural maps [6], non-linear measures [7], and voice source-related properties [8]. Those works used audio recordings from a single session. However, speech and voice disorders manifest themselves with varying severity on a day to day basis and that variability cannot be fully captured in a single session. Diagnosis and assessment of these disorders would benefit from unobtrusively monitoring and evaluating vocal features, from daily-life interactions. Conversational agents offer the potential of triggering automated assessment in this regard. Existing work has attempted to tackle this problem and through an ambulatory device [9], however this approach still requires employing an accelerometer on the neck surface to estimate glottal airflow.

One way in which speech and language therapists assess dysarthria is through auditory perceptual ratings of speech across the different speech subsystems of respiration, phonation, resonance, articulation and prosody. The phonatory aspects of speech (i.e. voice) can be assessed using the GRBAS scale [[10], in which phonation is described in terms of Grade (overall severity of dysphonia), Roughness, Breathiness, Asthenia and Strain. There have been a few attempts to match computational features to the GRBAS perceptual assessment. Jalalinajafabadi et al. used objective features extracted with the GPSP and Praat softwares to predict the GRBAS scale using Multiple Linear Regression and K-Nearest-Neighbour-Regression [11]. Saenz et al. also predicted the GRBAS attributes from MFFC using Learning Vector Quantization and a K Nearest Neighbour (KNN) classifier [12].

There are, to our knowledge, no studies on monitoring speech from disordered speech from unconstrained data collected in the field. The increasing popularity of commercial conversational agents such as Amazon Alexa or Google Home creates a unique opportunity to unobtrusively collect and monitor speech for populations with cognitive impairments. However, processing this data presents new challenges due to the unconstrained nature of the interactions and environmental factors, such as background noise and the need to process statements which may not be under the control of clinicians.

In this work, we identify acoustic features that contribute to different components of perceptual speech assessment, focusing particularly on phonatory aspects of speech. We present here a detailed case study of long-term speech collected by Alexa for a participants with significant speech impairments.

2. Material and Methods

2.1. Dataset

The UK-DRI CR&T has deployed Alexa voice assistants in the household of 14 PLwD as a part of a larger home study. An original 'check-in' app for PLwD to perform a daily health interview was developed as a part of the study. The app asks participants questions about how they feel (e.g. agitated, anxious or worried), how they slept and their upcoming plans for the day. The household inhabitants are also free to engage in unconstrained interactions with Alexa and use it for entertainment, information seeking, etc. In our study, a household is usually composed of a PLwD and carer (usually a spouse). They occasionally receive visitors, but visitor data was not assessed in

10.21437/Interspeech.2023-1729

this work. The study received ethical approval from the Surrey Borders Research Ethics Committee.

Data was pre-screened and any duplicate or recording that didn't contain speech directed towards Alexa was discarded. The full dataset consists of 12k audio recordings from English speakers (7046 from PLwD) collected over 6 months between 2021 and 2022. This dataset can be shared as extracted features upon reasonable request.

In this case study, we focus on one household, which had 1272 audio recordings, including 174 from the carer, 1017 from the PLwD and 81 from visitors. This data was acquired between September 8th 2021 and July 29th 2022. Although Alexa provides a transcript of the interaction, it was re-transcribed by our team as speech-to-text was often misunderstood: we found that for this participant the transcription was wrong 36.67% of the time.

Alexa provides us with audio recordings of a duration of maximum 8 s. For the purpose of this analysis, we will only considers recording where only the PLwD is talking. Each recording was manually annotated with several labels: who is speaking, topic of the interaction, speech assessment (respiration, phonation, resonance, articulation and prosody) and dysfluencies (hesitations, interjections, word/sentence repetitions). The labels have binary values to indicate absence/presence of an anomaly, except phonation which is rated between 0 and 3 for severity. Labelling the data presented some challenges, notably due to short speech samples (especially for features such as respiration), which is why for this exploratory study the presence or absence of a feature has been recorded rather than attempting a more robust or in-depth description.

The second author, a speech and language therapist, trained the first author in perceptual speech assessment. The first author annotated all the data and the second author annotated 10% of it for verification purposes. Overall inter-rater reliability was 88%, with 73% for the phonation severity score, 82% for breathiness, 90% for roughness, 83% for asthenia and 83% for strain. For each file, we extract 65 openSMILE features (version 2.4.1) [13, 14], using the low level Compare 16 feature set [15] (See [16] for a more detailed description of the features). Each recording is therefore divided into 60 ms chunks and features extracted for each chunk, we aggregate those features by computing the average and standard deviation for each file. Our feature vector is the concatenation of all the averages and standard deviations and therefore has dimension 130. Features are normalised between 0 and 1.

All source code required for conducting experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes.

2.2. Participant Information

Our participant was 83 years old, diagnosed with mixed Alzheimer's and vascular type dementia in 2016. She additionally had a background of chronic hypertension, hypothyroidism, hypercholesterolaemia, pre-diabetes, diverticulosis, and malignancy of the colon. During the course of this study she was also suffering from lung cancer. From the autumn of 2021 she was suffering with variable dysphonia likely as a result of tumour compression of the recurrent laryngeal nerve.

2.3. Labels

The audio recordings of the PLwD were annotated with the following labels for speech assessment:

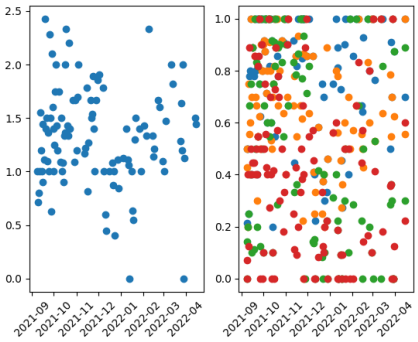

Figure 1: *Daily average of perceptual speech assessment. Left: in blue: phonation. Right: in blue breathiness, orange: roughness, red: asthenia, green: strain. The higher the value, the higher the severity of impairment.*

- **Respiration (5.5% of files)**: signs indicating possible reduced respiratory support for speech, e.g. shorter utterances per breath/additional breaths taken mid-phrase, lower volume or increasing dysphonia at the end of each breath.
- **Phonation (88.5% of files)**: Rated between 0 and 3 by level of severity (unimpaired, mildly impaired, moderately impaired, severely impaired). In addition, the presence or absence of the following phonatory features was recorded: *breathiness* (73.4% of files), *roughness* (64.5% of files), *asthenia* (49.8% of files) and *strain* (37.2% of files)
- **Resonance (2.0% of files)**: presence of hypernasality (excessive nasal resonance) or hyponasality (reduced nasal resonance).
- **Articulation (5.1% of files)**: presence of articulatory errors or disordered sounding articulation, e.g. imprecise articulation or perceptually excessively effortful articulation.
- **Prosody (5.1% of files)**: presence of unusual speech rhythm, melody, intonation, or stress patterns.
- **Language error (1.5% of files)**: presence of an apparent word or syntax error
- **Hesitation (12.5% of files)**: pauses of varying length that occur mid-utterance.
- **Disfluencies**:
  - *interjection (5.1% of files)*: extra sounds, syllables, or words often when thinking about what to say. Common interjections include: uh, um, well, like, you know, etc.
  - *word repetitions (0.8% of files)*: e.g. I (pause) I want.
  - *Partial or full phrase repetitions(0.6% of files)*: e.g. I like (pause) I like my new teacher.
  - *revisions (0.9% of files)*: e.g. I like cake (pause) cookies.

In this paper, for the sake of space, we will focus on the phonation labels, i.e overall phonation score, breathiness, roughness, asthenia and strain (See Fig. 1).

2.4. Classification Network

Our classification network is implemented with Keras and is composed of four fully connected layers with relu activation, 1024, 256, 64 and 16 neurons each and l2 regularisation (0.001), each followed by a Dropout layer (0.1), followed by a classification layer with sigmoid activation (softmax for the phonation component). Imbalance is addressed by setting appropriate class weights. We employ the binary cross-entropy loss. We train each model for a hundred epochs with early stopping cri-

teria of 5 and perform 5-fold cross-validation. We report the average F1-score as our evaluation metric.

3. Results

In this section, we present classification results, using all the available acoustic features, and after feature selection. Finally we discuss feature importance.

3.1. Classification

Our classification network is able to recognise phonation anomalies with a F1-score of 0.52, breathiness with a F1-score of 0.74, roughness with a F1-score of 0.64, asthenia with a F1-score of 0.68, strain with a F1-score of 0.69.

3.2. Features Selection

Since we have a high number of features (65x2), we first identify which ones are relevant for each subjective assessment. We, therefore, run a Kruskal-Wallis statistical test on each acoustic feature, to identify which ones are significantly different between our classes. We set the significance threshold to 0.001. This analysis identified 14 significantly different features for breathiness, 10 for roughness, 29 for asthenia, 57 for strain and 56 for phonation. Applying step backward feature selection leads to keeping 9 features for breathiness, 5 for roughness, 19 for asthenia, 57 for strain and 39 for phonation. After feature selection, we can recognise phonation anomalies with a F1-score of 0.55, breathiness with a F1-score of 0.71, roughness with a F1-score of 0.60, asthenia with a F1-score of 0.65, strain with a F1-score of 0.74.

3.3. Feature Importance

To evaluate feature importance, we evaluate performance of the trained network with each feature column replaced, one by one, by Gaussian noise and observe how it fares compared to the best performance. This was also evaluated using 5-fold cross-validation.

3.3.1. Phonation

The features contributing to the Phonation component are the F0final, jitterLocal, shimmerLocal, logHNR, audspec_lengthL1norm, pcm_RMSenergy, audSpec_Rfilt (components 0, 4, 5, 6, 7, 16, 22, 24, 25), pcm_fftMag_spectralRollOff25.0, pcm_fftMag_spectralRollOff75.0, pcm_fftMag_spectralFlux, pcm_fftMag_psySharpness, pcm_fftMag_spectralCentroid, pcm_fftMag_spectralEntropy, mfcc (components 1, 3, 4, 5, 6, 7, 8, 9, 10, 14). The most important feature are the pcm_fftMag_spectralRollOff75.0 and the 4th component of mfcc (See Fig. 2).

3.3.2. Breathiness

The features contributing to breathiness are the 6th, 7th, 15th and 16th components of audSpec_Rfilt, and the 4th, 5th and 7th components of mfcc. The most important feature in evaluating breathiness is the 7th component of mfcc (See Fig. 3).

3.3.3. Roughness

Features contributing to roughness are the 4th, 5th and 7th components of audSpec_Rfilt, and audspecRasta_lengthL1norm.

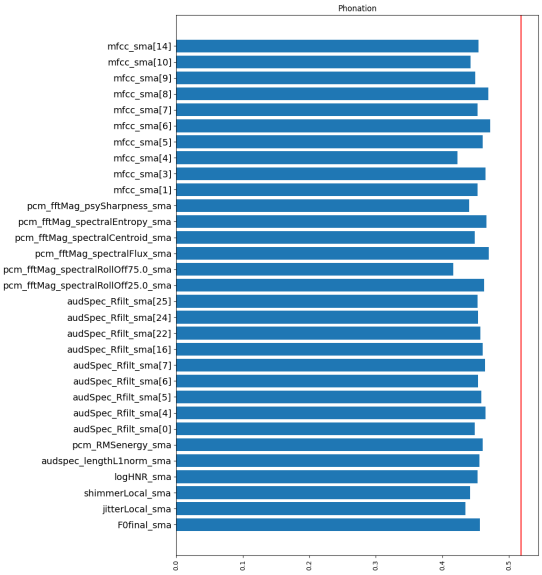

Figure 2: *Performance (F1-score) when classifying phonation and replacing each feature by Gaussian noise, one by one.*

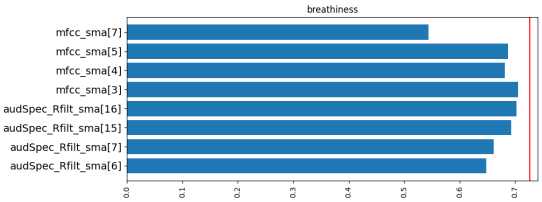

Figure 3: *Performance (F1-score) when classifying breathiness and replacing each feature by Gaussian noise, one by one.*

The most important feature are the 4th and 5th components of audSpec_Rfilt (See Fig. 4).

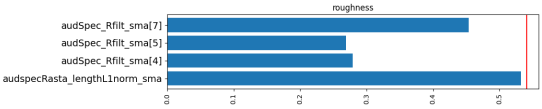

Figure 4: *Performance (F1-score) when classifying roughness and replacing each feature by Gaussian noise, one by one.*

3.3.4. Asthenia

Features contributing to asthenia are the audSpec_Rfilt (components 5, 7, 8, 10, 11, 14, 15, 16, 17, 18), 3rd, 4th, 7th, 9th and 10th components of mfcc, and pcm_fftMag_spectralFlux. The most important features in identifying asthenia is the 9th component of mfcc (See Fig. 5).

3.3.5. Strain

Features contributing to the strain are F0final, voicingFinalUnclipped, jitterLocal, jitterDDP, shimmerLocal, audspec_lengthL1norm, pcm_RMSenergy, audSpec_Rfilt (components 0, 1, 20, 21, 22, 23, 24, 25), pcm_fftMag_fband1000-4000, pcm_fftMag_spectralRollOff[25.0, 50.0, 75.0, 90.0], pcm_fftMag_spectralFlux, pcm_fftMag_spectralCentroid, pcm_fftMag_spectralEntropy, pcm_fftMag_psySharpness, pcm_zcr, mfcc (components 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14). The most important feature are the pcm_fftMag_spectralRollOff75.0 and the 13th component of mfcc (See Fig. 6).

F0final	The smoothed fundamental frequency contour
voicingFinalUnclipped	The voicing probability of the final fundamental frequency candidate. Unclipped means, that it was not set to zero when is falls below the voicing threshold
jitterLocal	The local (frame-to-frame) Jitter (pitch period length deviations)
jitterDDP	The differential frame-to-frame Jitter (the 'Jitter of the Jitter')
shimmerLocal	The local (frame-to-frame) Shimmer (amplitude deviations between pitch pe- riods)
logHNR	log harmonics-to-noise ratio
audspecRasta_lengthL1norm	Relative Spectral Transform applied to Auditory Spectrum and the magnitude of the L1 norm
audspec_lengthL1norm	magnitude of L1 norm of the auditory spectrum
audSpec_Rfilt	Relative Spectral Transform (RASTA)-style filtered applied to Auditory Spectrum
pcm_RMSenergy	Root-mean-square signal frame energy
pcm_fftMag_fband1000-4000	fft magnitude of this frequency band
pcm_fftMag_spectralFlux	spectral flux of the magnitude of the FFT
pcm_fftMag_spectralRollOff	spectral roll-off points of the magnitude of the FFT
pcm_fftMag_spectralCentroid	spectral centroid of the magnitude of the FFT
pcm_fftMag_spectralEntropy	spectral entropy of the magnitude of the FFT
pcm_fftMag_psySharpness	Psychoacoustic sharpness of the magnitude of the FFT
mfcc	Mel-Frequency cepstral coefficients
pcm_zcr	Zero-crossing rate of time signal (frame-based)

Table 1: *Explanation of the openSMILE features selected in this paper*

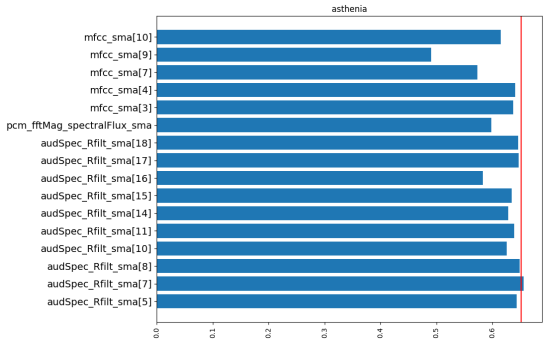

Figure 5: *Performance (F1-score) when classifying asthenia and replacing each feature by Gaussian noise, one by one.*

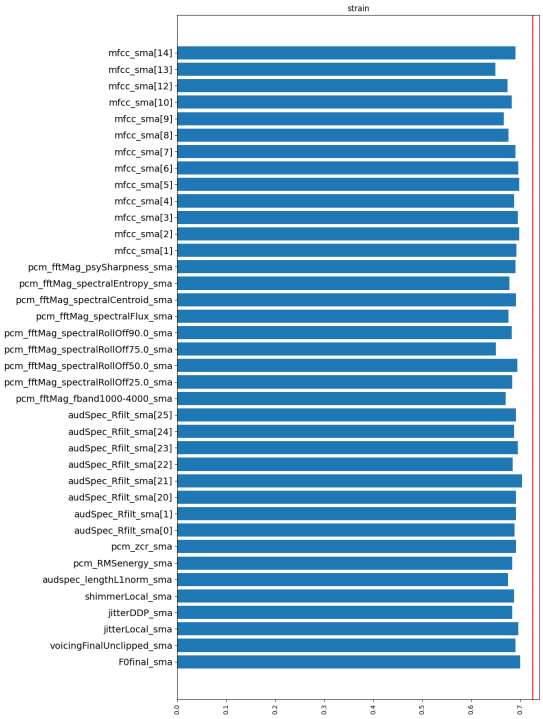

Figure 6: *Performance (F1-score) when classifying strain and replacing each feature by Gaussian noise, one by one.*

4. Conclusions

In this work, we perform a feature importance analysis to match perceptual speech assessment to acoustic features for phonation ratings. Isolating those acoustic features will allow us to devise a speech quality index to automate speech assessment. Performance is lower for the Phonation rating, this can probably be explained by the difficulty to properly label severity on a 4-point scale. Indeed, when considering a binary label (absence/presence of phonation anomaly), the network performs much better: F1-score of 0.86 on the 5-fold cross-validation.

This work focuses on a case study of one participant with a lot of recordings and variable speech quality. In future works, we will continue our longitudinal analysis of this data. We hope that this research will pave the way towards automatic speech assessment. Since we have now isolated relevant features for each scale, we will also develop a computational voice quality index. Limitations of this work are related to the unconstrained nature of the data, i.e. short utterances and sometimes noisy environment. This work is the first stepping stone to automating speech assessment and designing a monitoring system for speech impairments and subsequently for cognitive decline. It also explores the possibility of fully exploiting the opportunity created by voice assistants.

5. Acknowledgements

This work is supported by the UK Dementia Research Institute (UKDRI-7003) which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

This work would not be possible without the infrastructure for data collection established by our institute. We also express our warmest sentiments in memory of our participant, whose patience and dedication graced not only this work, but enriched all of those motivated to support technology development in this field.

6. References

[1] K. E. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task," *Neurological sciences*, vol. 26, pp. 243–254, 2005.

[2] S. Ahmed, C. A. de Jager, A.-M. Haigh, and P. Garrard, "Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed alzheimer's disease." *Neuropsychology*, vol. 27, no. 1, p. 79, 2013.

[3] V. Taler and N. A. Phillips, "Language performance in alzheimer's disease and mild cognitive impairment: a comparative review," *Journal of clinical and experimental neuropsychology*, vol. 30, no. 5, pp. 501–556, 2008.

[4] A. Oulhaj, G. K. Wilcock, A. D. Smith, and C. A. De Jager, "Predicting the time of conversion to mci in the elderly: role of verbal expression and learning," *Neurology*, vol. 73, no. 18, pp. 1436–1442, 2009.

[5] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. Meilán, "Speech biomarkers of risk factors for vascular dementia in people with mild cognitive impairment," *Frontiers in Human Neuroscience*, 2022.

[6] S. Hadjitodorov, B. Boyanov, and B. Teston, "Laryngeal pathology detection by means of class-specific neural maps," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 1, pp. 68–73, 2000.

[7] M. Little, P. Mcsharry, S. Roberts, D. Costello, and I. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Nature Precedings*, pp. 1–1, 2007.

[8] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of speech, language, and hearing research*, vol. 43, no. 2, pp. 469–485, 2000.

[9] D. D. Mehta, J. H. Van Stan, M. Zañartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortés, H. A. Cheyne, and R. E. Hillman, "Using ambulatory voice monitoring to investigate common voice disorders: Research update," *Frontiers in bioengineering and biotechnology*, vol. 3, p. 155, 2015.

[10] M. Hirano and K. R. McCormick, "Clinical examination of voice by minoru hirano," 1986.

[11] F. Jalalinajafabadi, *Computerised GRBAS assessment of voice quality*. The University of Manchester (United Kingdom), 2016.

[12] N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006.

[13] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[15] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, vol. 8. ISCA, 2016, pp. 2001–2005.

[16] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.