IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Prosodic feature selection for automatic quality assessment of oral productions in people with Down syndrome

David Escudero-Mancebo, Mario Corrales-Astorgano,
Valentín Cardeñoso-Payo, César González-Ferreras

Department of Computer Science
University of Valladolid
descuder@infor.uva.es

Abstract

Evaluation of prosodic quality is always a challenging task due to the nature of prosody with multiple form-function valid profiles. When voice of people with Down syndrome (DS) is analyzed, diversity increases making the problem even more challenging. This work is framed in our activities for developing learning games for training oral communications of people with intellectual disabilities. In this context automatic evaluation of prosodic quality is a must for deciding whether game users should repeat activities or continue playing and to inform therapists about the particular difficulties of users. In this paper we present a procedure for the selection of informative prosodic features based on both the distance between human rated right and wrong productions and the distance with respect to productions of typical users. A main contribution with respect to previous works stems from the use of mixed models to rate the impact of the type of activity and speaker dependence when estimating the quality of the prosodic productions.

Index Terms: Down syndrome speech, Computer Assisted Pronunciation Assessment.

1. Introduction

Prosody is an important component of speech communication because it is responsible for fundamental functions such as grouping linguistic units, pausing, word accent and sentence purpose (declarative, interrogative, exclamatory or imperative) and also other higher level functions like emotions and pragmatics [1]. The low control of prosody or its inappropriate production can stigmatize speakers and limit their options to get integrated in society [2]. Such could be the case of people with intellectual disabilities in general and speakers with Down syndrome (DS) in particular, which is a population characterized by special needs on language control and prosodic production (with notable exceptions) [3, 4, 5]. As far as prosody is concerned, Kent and Vorperian [4] report disfluencies (stuttering and cluttering) and impairments in the perception, imitation and spontaneous production of prosodic features; while Heselwood et al. [6] have connected some of the speech errors with difficulties in the identification of boundaries between words and sentences. In previous work we empirically showed the clear contrast between the voice of speakers with Down syndrome and typical speakers by performing perceptual and automatic identification tests from signal [7].

This work has been partially funded by Ministerio de Economía, Industria y Competitividad and the European Regional Development Fund FEDER (project TIN2017-88858-C2-1-R) and by Junta de Castilla y León (project VA050G18).

In [8] we took a step forward to analyze the possibilities to assess DS voice oral productions quality by using a similar procedure. While the problem of DS voice identification reached more than 90% accuracy with an SVM classifier [7], the problem of quality assessment reached only about 78.5% using the same type of classifier and the same training feature set. In this paper we analyze the training data used in the mentioned previous works to understand the reasons for this classification performance differences at the time that results give cues for exploring different paradigms in future works.

Software tools and learning games have been devised for intellectual disabled people to train specific competences [9, 10, 11, 12]. There are methods that voice therapists employ with speakers with specific speech problems [13]. Some of these methods, partially, have been implemented as software tools that help therapists to work with their patients or allow patients to carry out complementary exercises in an autonomous way [14]. In [15] we presented a tool to train prosody and pragmatics for speakers with DS. A set of perceptual and production activities are interleaved in a graphic adventure video game with an adapted interface that takes into account the special characteristics of individuals with Down syndrome: poor short term memory [16], attention deficits [17], problems to integrate information and deficits of language development [18]. So far, the video game has been used successfully with real users, with the assistance of an adult (the teacher, the therapist or a relative). The use of the tool has allowed the recording of a speech corpus of people with Down syndrome. The final goal of the research presented in this paper is the analysis of the potential of these recordings to train an automatic assessment system. In the medium term, this system will be integrated into the video game to allow users to train autonomously. A complete description of oral activities can be found in [19].

There are several works on automatic assessment of speech quality in atypical voices described in the literature [20, 21, 22]. However, in computer assisted pronunciation training, not only assessment is important, but also reporting information about the reasons that led the system or the expert to judge a given utterance as correctly or incorrectly produced. In [23] authors analyze how different components of speech production impact speech intelligibility in DS. In this paper we systematically analyze the prosodic features of the utterances of the corpus in order to select the most informative features and their values for predicting oral productions quality. To perform this analysis, we triangulate information obtained from the distances between right and wrong DS productions (at the glance of human evaluators) with information obtained from distances between DS productions and productions of typical users. We then impose the requirements of separation between groups and consistency.

10.21437/IberSPEECH.2021-1

In a second step, by using logistic mixed regression models, we find that features related to temporal domain are more efficient for this task than other prosodic features related to $F0$ or energy. In addition, we provide evidences that it is important to consider not only the speaker (already shown in [8]) but also the particular training activity for improving the accuracy of the automatic assessment system.

The structure of the paper is as follows. The experimental procedure section details the corpus compiled and the manual evaluation of the utterances. The procedure for the individual analysis of the different prosodic features is also presented. The results section lists the selected prosodic features and its capabilities for modeling the quality of the utterances taking into account human scores. We end the paper with a discussion that includes limitations, future work and conclusions.

2. Experimental procedure

2.1. Corpus recording

The corpus was recorded by using a graphic adventure video game [15] that requires users to perform a set of activities related to prosodic perception and production skills in order to continue playing. All the oral productions and user interactions are recorded and classified per activity and speaker while the user is playing. The current corpus has been compiled in different sessions. It has recordings of 23 speakers with Down syndrome, 966 utterances of 40 different production activities, which is about 1 hour and 10 minutes duration. More details about gender, age and metal capabilities can be found in [8]. From this work we select those 5 speakers with more than 40 utterances each (606 utterances in total), with the aim of obtaining representative results. We call this corpus θ_{DS}.

We used a second corpus of typical speakers to analyze voice of speakers with Down syndrome when contrasted with voice of typical speakers. This corpus contains a subset of the sentences recorded by the speakers with Down syndrome and has recordings of 22 speakers, with 250 utterances in total. We call this corpus θ_{TS} (details about gender and age can be found in that was used in [7])

The recordings of both corpora were made at 44,100 Hz with a Logitech PC Headset 960 USB microphone.

2.2. Human based quality evaluation

Two complementary human based evaluations have been performed:

Real-time evaluation: The production activities are assessed in real time by a therapist, seated next to the player with a secondary keyboard. The therapist can evaluate the activity as Right (θ_R) or Poor (θ_P) to let the player continue, depending of the production quality, or as Wrong (θ_W) to ask him/her to repeat the utterance. The therapist is responsible to assess the production of the gamer and the consequence is that the player has to try again to pronounce correctly until the therapist considers that he or she can continue playing. The real time scores divide the corpus: $\theta_{DS} = \theta_R \cup \theta_W \cup \theta_P$.

Off-line evaluation: Each of the utterances was rated off-line by an expert in prosody who participated in the design of the video game. The judgments were binary, corresponding to the decision of whether the utterance should be repeated or not. She declared to use the following decision criteria, adapting them to the specific activity proposed: adjustment to the expected modality (intonation); preservation of the difference between lexical stress (stressed vs. unstressed syllables) and accent (accented vs. unaccented syllables); and adjustment to the organization in prosodic groups and distinction between function and content words (phrasing). The off-line scores divide the corpus $\theta_{DS} = \theta_{R'} \cup \theta_{W'}$; R' indicating right and W' indicating wrong productions.

Both evaluations have been compared leading to consistency rates going from 79.4% to 64.1% depending on the speaker (doing R and P assignments equivalent to R')

2.3. Processing and selection of prosodic features

The openSmile toolkit [24] was used to extract acoustic features from each recording of the corpus. The GeMAPS feature set [25] was selected due to the variety of acoustic and prosodic features contained in this set: frequency related features, energy related features and temporal features. The arithmetic mean and the coefficient of variation along the utterance were calculated on these features. Furthermore, 4 additional temporal features were added: the silence and sounding percentages, silences per second and the length mean of silences. These last 4 features were calculated using the silences and sounding intervals generated by Praat software [26], which uses an intensity threshold, a minimum silent interval duration and a minimum sounding interval duration to identify these intervals (Praat default values were used). In total, 92 features were used: 10 from frequency domain, 10 from energy domain, 11 from temporal domain and 61 from spectral domain. The complete description of these features can be found in [7].

As selection criteria we require the feature f to satisfy:

1. **Separation:** there must be statistical significant differences between the values of f in the groups θ_R and θ_W (Mann-Whitney test with p-value$<$0.01) which implies that clear differences between right and wrong utterance are observed.
2. **Consistency:** being f_T, f_R and f_W the mean value of the feature f in the groups θ_{TS} (typical speakers), θ_R and θ_W respectively, it must be satisfied that $|f_T - f_R| < |f_T - f_W|$ which implies that right utterances are closer than wrong utterances to the typical ones.

We apply this procedure with both real-time and off-line evaluations. In the case of real-time evaluation the procedure is repeated for every pair of groups. The comparison of results obtained with both evaluations permits to discuss about the possible reasons for disagreement.

2.4. Analysis of the impact of features on quality

Logistic mixed effects regression models were used to measure the impact of speaker and activity on the automatic assessment of quality. This regression model takes into account that the I observations came from A different activities and S different speakers. The full model is given by

$$\begin{aligned} y_{i,a,s} = \beta_0 + A_{0,a} + S_{0,s} + \\ (\beta_\delta + A_{\delta,a} + S_{\delta,s}) X_{i,a,s} + \epsilon_{i,a,s} \end{aligned} \quad (1)$$

where β_0 is the fixed intercept and $A_{0,a}$ and $S_{0,s}$ are the random intercepts introduced by activity and speaker respectively; $A_{\delta,a}$, and $S_{\delta,s}$ are the random slopes to be added to the fixed slope β_δ. The most informative acoustic features are the fixed effect and speaker and activity are the random effects. A binomial distribution for y was used in order to build the logistic regression models. Different configurations of the mixed model

Table 1: *List of the automatically selected frequency, energy and temporal features. All the features in columns Off-line evaluation have statistically significant differences (Mann-Whitney test with p-value<0.01) between right and wrong productions of speakers with Down syndrome. The asterisk in Real-time evaluation columns means statistically significant differences (Mann-Whitney test with p-value<0.01) between first and third column (placed in the first column), between first and second column (when placed in the second column) or between the second and third column (when placed in the third column). The meaning of the features can be seen in [7]. In cells we present 95% confidence interval of the mean value. The units are reported in [24].*

			Off-line evaluation		**Real time evaluation**		
		Typical speakers	DS Right productions	DS Wrong productions	DS Right cont. prod	DS Wrong cont. prod	DS Poor productions
F0 domain							
f1	F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	(2.40, 2.88)	(1.91, 2.67)	(2.73, 3.66)	(1.37, 2.23)*	(2.08, 3.03)	(3.48, 4.74)*
f2	jitterLocal_sma3nz_stddevNorm	(1.11, 1.21)	(1.32, 1.43)	(1.52, 1.66)	(1.35, 1.48)	(1.39, 1.56)	(1.40, 1.55)
Energy domain							
e1	loudness_sma3_percentile20.0	(0.91, 1.01)	(0.71, 0.78)	(0.63, 0.71)	(0.65, 0.72)	(0.72, 0.84)	(0.67, 0.77)
Temporal domain							
d1	loudnessPeaksPerSec	(5.64, 5.89)	(4.10, 4.32)	(3.76, 4.05)	(4.23, 4.49)*	(3.80, 4.14)*	(3.61, 3.92)
d2	StddevVoicedSegmentLengthSec	(0.14, 0.16)	(0.18, 0.23)	(0.25, 0.33)	(0.20, 0.26)	(0.20, 0.28)	(0.19, 0.27)
d3	soundingPercentage	(0.88, 0.91)	(0.89, 0.91)	(0.73, 0.79)	(0.88, 0.91)*	(0.82, 0.88)	(0.74, 0.81)*
d4	silencesPerSecond	(0.35, 0.44)	(0.28, 0.35)	(0.52, 0.63)	(0.35, 0.44)*	(0.36, 0.50)	(0.45, 0.58)
d5	silencesMean	(0.14, 0.19)	(0.13, 0.18)	(0.31, 0.41)	(0.14, 0.19)*	(0.18, 0.27)	(0.28, 0.41)

were compared in terms of the modeling capabilities with the use of the `lme4` package [27].

A reduced number of variables is used for the algorithm to iterate with 606 points, at most 3 fixed effects and 2 random effects. This procedure permits to assess both the relative importance of the acoustic features (fixed factors) and speaker and activity (random factors) on the perceived quality. We select the most informative feature of each of the three domains as fixed factors.

3. Results

Table 1 presents the 95% confidence interval of the mean value of the selected features separated by groups. Only 8 out of the 92 analyzed input features satisfy the established criteria when off-line evaluation data are contrasted (see table 1) (27 features were selected in [7] and 21 in [8]). Only 5 of them do when real-time evaluation is contrasted (asterisks in the last three columns of table 1). In fact, in real-time evaluation, when the groups DS Right vs. DS Wrong and DS Wrong vs. DS Poor are compared, only one and two variables respectively satisfy the imposed conditions (features f1, e1 and d3). This result suggests a richer evaluation in off-line conditions with more features in all the domains. In real-time evaluation the values of the energy domain variables are inconsistent in what concerns to group separation: no significant differences between groups and the closer to the typical values does not imply the more quality. Results suggest that recordings were evaluated by Poor when abnormal values of the temporal domain or F0 domain were observed (features f1, d1, d3-d5) and that utterances were marked as Wrong when speed was low (d1 feature).

Concerning the selected features, F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2 represents the range of 20-th to 80-th of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz and jitterLocal_sma3nz_stddevNorm represents the coefficient of variation of the deviations in individual consecutive F0 period lengths. In energy domain, loudness_sma3_percentile20.0 means the percentile 20-th of estimate of perceived signal intensity from an auditory spectrum. Finally, related with the temporal domain, loudnessPeaksPerSec means the number of the loudness peaks per second and StddevVoicedSegmentLengthSec represents the standard deviation of continuously voiced regions. soundingPercentage represents the duration percentage of voiced regions, silencesPerSecond means the number of silences per second and silencesMean represents the length mean of unvoiced regions.

The values of the confidence intervals in table 1 show that the wider the F0 range (higher f1 feature) and the less stable f0 contour (higher f2 feature) the more abnormal the utterance is perceived; the weaker the intensity (lower e1 feature) the more penalized the utterance is; utterances belonging to the wrong and poor groups are slower (lower d1 feature), have more speed changes (higher d2 feature), with more inner pauses (lower d3 and higher d4 feature) or longer pauses (higher d5 feature).

Table 1 also permits to contrast the values of features in typical vs. DS speakers. Focusing on off-line evaluation, we observe that the gap between Typical and Right productions is relevant for features f2, e1, d1, d2, d4 (with not overlapping intervals). The distance between Typical and Right utterances is higher than the distance between Right and Wrong utterances for features f2, e1 and d1. f1 and d5 features present the most overlapped intervals between Right and Typical utterances.

Table 2 shows how accurate a set of logistic regression models represent the working data. We select a feature per domain as the features in the same domain exhibit a high correlation. The incremental inclusion of new variables in the ANOVA test permits to show that all the variables significantly contribute in the modeling. The use of the variable related to duration domain (Dur in m3 and m10 models) and the inclusion of the activity in the model (m7 model), offer the most significant modelling improvements (more AIC and deviance reduction and more Acc increase).The use of random factors is a need for improving the quality of the modelling (AIC goes from 716.19 to 612.60 and Acc from 70% to 80%). Slopes of the random factors do not contribute to improve the modeling (m11 and m12 rows).

The feature selection procedure allows to identify the 8 most informative variables to predict prosodic quality from the 92 analyzed variables with satisfactory results: an automatic classifier SVM trained with the 92 variables performs with 72.0% accuracy and with the 8 variables selected performs with 71.0% (SMO -Sequential Minimal Optimization- implementation of Weka tools using normalized poly kernel with exponent

Table 2: *Summary of the sequential ANOVA test of the mixed logistic regression models when the quality of the utterance is predicted using the binary off-line evaluation. $F0$ is the variable f1 in table 1, En is the variable e1, Dur is the variable d5, $Speaker$ ranges from 1 to 5, $Activity$ ranges from 1 to 40. I means intercept and S refers to both slope and intercept.The quality metrics are the ones reported by the command anova of the package lme4 and Acc is the accuracy of the prediction of the training samples. Sig. codes: 0 '***' 0.001 '**' 0.01 '*'. AIC is the Akaike's Information Criterion [28].*

	Fixed effect			Random factor		Quality of the model					
Model	F0	En	Dur	Speaker	Activity	AIC	deviance	Chisq	Pr(>Chisq)	sig	Acc
m1	X					775.76	771.76				65%
m2		X				774.68	770.68	1.083			66%
m3			X			723.98	719.98	50.700			68%
m4	X	X				720.82	714.82	5.159	0.02313	*	69%
m5	X	X	X			716.19	708.90	5.921	0.01496	*	70%
m6	X			I		761.36	755.36				66%
m7	X				I	684.19	678.19	77.1729	$<$ 2.2e-16	***	74%
m8	X			I	I	666.17	658.17	20.0255	7.642e-06	***	75%
m9	X	X		I	I	654.38	644.38	13.7854	0.0002049	***	77%
m10	X	X	X	I	I	612.60	600.60	43.7799	3.675e-11	***	80%
m11	X	X	X	I	S	614.31	600.31	0.2872	0.5920068		80%
m12	X	X	X	S	S	616.31	600.31	0.0000	1.0000000		80%

2 [29]). Including speaker and activity information, the SVM classifier (same configuration) predicts prosodic quality with 76.8% accuracy with 8 variables and 76.7% with the 92 acoustic features (exponent 1 poly kernel in the last case).

4. Discussion

The acoustic features belonging to the temporal domain seem to be effective for the assessment of oral turns independently of speaker and activity: 5 variables out of the 8 selected features refer to temporal features (table 1) and the models including the duration of the pauses are the best predicting the quality of the utterances (table 2). This is an attractive result because the computation of this type of features, in contrast with the ones belonging to the F0 or spectral domain, is more robust in the face of the adverse conditions that could occur with users with DS. In general, the pitch detection algorithms produce more errors in pathological voices than in typical voices [30]. Furthermore, features related to the temporal domain can be easily related to disfluent speech (stuttering or cluttering) that, although not universal, is a common problem of this population [31, 32, 33].

The mixed regression model indicates that not only considering the speaker is important (already shown in [8]) but also the particular activity performed by the speaker during the recording. The specification of the particular activity was not relevant to identify whether the oral turn corresponded to a speaker with Down syndrome or not in previous works [7]. Nevertheless, here it appears to be relevant for assessing the quality of the oral turns (significantly higher precision in models m7-m10).

The search of new features, or combinations of the ones already computed, that could be related to the activity or type of activity is proposed as future work in order to improve the results, at the time that a bigger corpus is compiled for testing more sophisticated models or alternative machine learning techniques. This is a challenging task because the video game has diverse activities for players to train different language functions like asking, expressing opinions, social interaction... prosodic functions like chunking or prominence in different production modes: reading, elicited or free speech. It is present work the use of the utterances of the corpus for compiling an unsupervised classification of the activities that takes into account the different acoustic prosodic features and human judgments of quality.

5. Conclusions

The paper has presented a feature selection procedure that profits evidences from different human based evaluations and that benefits as well of empirical observations that concern with differences between Down syndrome utterances with respect to typical speakers ones.

The procedure has shown to be efficient so that 8 selected features permit predicting prosodic quality as accurately as 93 features ensemble. It permits identifying the most discriminant features per domain: temporal, frequency and energy related domain.

It has been shown and discussed the reasons why improving the accuracy of the classifier requires the consideration of the type of activity and specific profile of the user of the training tool. This fact highlights the need to implement specialized classifiers on the different type of activities and the implementation as well of user adaption techniques in future work.

6. References

[1] P. Roach, *English phonetics and phonology fourth edition: A practical course*. Ernst Klett Sprachen, 2010.

[2] B. Wells, S. Peppé, and M. Vance, "Linguistic assessment of prosody," *Linguistics in clinical practice*, pp. 234–265, 1995.

[3] R. S. Chapman and L. Hesketh, "Language, cognition, and short-term memory in individuals with Down syndrome," *Down Syndrome Research and Practice*, vol. 7, no. 1, pp. 1–7, 2001.

[4] R. D. Kent and H. K. Vorperian, "Speech impairment in Down syndrome: A review," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 1, pp. 178–210, 2013.

[5] V. Stojanovik, "Prosodic deficits in children with Down syndrome," *Journal of Neurolinguistics*, vol. 24, no. 2, pp. 145–155, 2011.

[6] B. Heselwood, M. Bray, and I. Crookston, "Juncture, rhythm and planning in the speech of an adult with down's syndrome," *Clinical Linguistics & Phonetics*, vol. 9, no. 2, pp. 121–137, 1995.

[7] M. Corrales-Astorgano, D. Escudero-Mancebo, and C. González-Ferreras, "Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome," *Speech Communication*, vol. 99, pp. 90 – 100, 2018.

[8] M. Corrales-Astorgano, P. Martínez-Castilla, D. Escudero-Mancebo, L. Aguilar, C. González-Ferreras, and V. Cardeñoso-Payo, "Automatic assessment of prosodic quality in Down syndrome: Analysis of the impact of speaker heterogeneity," *Applied Sciences*, vol. 9, no. 7, p. 1440, 2019.

[9] A. R. Cano, Á. J. García-Tejedor, C. Alonso-Fernández, and B. Fernández-Manjón, "Game analytics evidence-based evaluation of a learning game for intellectual disabled users," *IEEE Access*, vol. 7, pp. 123 820–123 829, 2019.

[10] L. E. García, R. J. Mejía, A. Salazar, and C. E. Gómez, "Un videojuego para estimular habilidades matemáticas en personas con síndrome de Down," *Revista ESPACIOS*, vol. 40, no. 05, 2019.

[11] K. Prena and J. L. Sherry, "Parental perspectives on video game genre preferences and motivations of children with Down syndrome," *Journal of Enabling Technologies*, vol. 12, no. 1, pp. 1–9, 2018.

[12] M. S. Del Rio Guerra, J. Martin-Gutierrez, R. Acevedo, and S. Salinas, "Hand gestures in virtual and augmented 3d environments for Down syndrome users," *Applied Sciences*, vol. 9, no. 13, p. 2641, 2019.

[13] D. R. Boone, S. C. McFarlane, S. L. Von Berg, and R. I. Zraick, *The voice and voice therapy*. Pearson/Allyn & Bacon Boston, 2005.

[14] W. R. Rodríguez, O. Saz, and E. Lleida, "A prelingual tool for the education of altered voices," *Speech Communication*, vol. 54, no. 5, pp. 583–600, 2012.

[15] C. González-Ferreras, D. Escudero-Mancebo, M. Corrales-Astorgano, L. Aguilar-Cuevas, and V. Flores-Lucas, "Engaging adolescents with Down syndrome in an educational video game," *International Journal of Human–Computer Interaction*, vol. 33, no. 9, pp. 693–712, 2017.

[16] R. Chapman and L. Hesketh, "Language, cognition, and short-term memory in individuals with Down syndrome," *Down Syndrome Research and Practice*, vol. 7, no. 1, pp. 1–7, 2001.

[17] M. H. Martínez, X. P. Duran, and J. N. Navarro, "Attention deficit disorder with or without hyperactivity or impulsivity in children with Down's syndrome," *International Medical Review on Down Syndrome*, vol. 15, no. 2, pp. 18–22, 2011.

[18] R. S. Chapman, "Language development in children and adolescents with Down syndrome," *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 3, no. 4, pp. 307–312, 1997.

[19] D. Escudero-Mancebo, M. Corrales-Astorgano, V. Cardeñoso-Payo, L. Aguilar, C. González-Ferreras, P. Martínez-Castilla, and V. Flores-Lucas, "Prautocal corpus: A corpus for the study of down syndrome prosodic aspects." *Languages Resources and Evaluation*, vol. Under review, 2021.

[20] D. Le and E. M. Provost, "Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation," in *INTERSPEECH*, 2014.

[21] M. Tu, V. Berisha, and J. Liss, "Interpretable objective assessment of dysarthric speech based on deep neural networks." in *INTERSPEECH*, 2017, pp. 1849–1853.

[22] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, and X. Zou, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Computer Speech & Language*, vol. 56, pp. 80–94, 2019.

[23] D. O'Leary, A. Lee, C. O'Toole, and F. Gibbon, "Perceptual and acoustic evaluation of speech production in Down syndrome: A case series," *Clinical linguistics & phonetics*, pp. 1–20, 2019.

[24] 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*. ACM, 2013, pp. 835–838.

[25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[26] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[28] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[29] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.

[30] S.-J. Jang, S.-H. Choi, H.-M. Kim, H.-S. Choi, and Y.-R. Yoon, "Evaluation of performance of several established pitch detection algorithms in pathological voices," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 620–623.

[31] J. Van Borsel and A. Vandermeulen, "Cluttering in Down syndrome," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 6, pp. 312–317, 2008.

[32] D. Devenny and W. Silverman, "Speech dysfluency and manual specialization in Down's syndrome," *Journal of Intellectual Disability Research*, vol. 34, no. 3, pp. 253–260, 1990.

[33] K. Eggers and S. Van Eerdenbrugh, "Speech disfluencies in children with Down Syndrome," *Journal of Communication Disorders*, 2017.