Suprasegmental Information Modelling for Autism Disorder Spectrum and Specific Language Impairment Classification

David Martínez[1], *Dayana Ribas*[2], *Eduardo Lleida*[1], *Alfonso Ortega*[1], *Antonio Miguel*[1]

[1]Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
[2]Advanced Technologies Application Center (CENATAV), La Habana, Cuba

david@unizar.es, dribas@cenatav.co.cu, (lleida | ortega | amiguel)@unizar.es

10.21437/Interspeech.2013-66

Abstract

This paper investigates the efficiency of several acoustic features in classifying pervasive developmental disorders, pervasive developmental disorders not-otherwise specified, dysphasia, and a group of control patients. One of the main characteristics of these disorders is the misuse and misrecognition of prosody in daily conversations. To capture this behaviour pitch, energy, and formants are modelled in long-term intervals, and the interval duration, shifted-delta cepstral coefficients, AM modulation index, and speaking rate complete our acoustic information. The concept of total variability space, or iVector space, is introduced as feature extractor for autism classification. This work is framed in the Interspeech 2013 Computational Paralinguistics Challenge as part of the Autism Subchallenge. Results are given on the Child Pathological Speech Database (CPSD), and an 87.6% and 45.1% unweighted average recall are obtained for the typicality (typical vs. atypical developing children) and diagnosis (classification into the 4 groups) tasks, respectively, on the development dataset. In addition, the combination of the new and the baseline features offers promising improvements.

Index Terms: Autism, Computational Paralinguistics, Challenge, Prosody, iVectors

1. Introduction

The term autism spectrum disorder (ASD) describes a range of disorders characterized by a triad of impairments: atypical development in reciprocal social interaction, atypical communication, and restricted, stereotyped and repetitive behaviours [1]. The range of disorders usually include autism disorder (AD), Asperger syndrome (AS), and pervasive developmental disorder not otherwise specified (PDD-NOS) [2]. Recent investigations indicate that the median of prevalence estimates of ASD is 62/10000 [3]. While different studies show variations, the evidences are not strong enough to conclude if there are differences in prevalence caused by geographic regions, ethnic, cultural, or socioeconomic factors.

On the other hand, specific language impairment (SLI), commonly interchanged with the term developmental dysphasia (DYS), is a speech impairment that affects the mastering of language skills, in particular structural aspects like phonology and syntax. In contrast, the abnormal use of pragmatics is the most evidence in ASD [4]. However, the boundaries of both diagnoses are unclear, and some authors prefer to speak of a continuum instead of completely independent groups [4].

As pointed out in several works [5, 6], there is a need for clinicians to objectively classifiy the different disorders of ASD and SLI. The correct identification requires a lot of expertise, and the decisions have a subjective component. Automatic classification tools made by computers would eliminate this problem. Given the communication problems shown in speech that people affected by ASD have, speech technologies seem a good option to help to achieve this goal. There are studies that focus on specific aspects of the speech that characterize ASD patients. For example, in [5] the authors represent prosodic features with a large set of statistical measurements of pitch and energy, and model it with static and dynamic classification algorithms. They use it to classifiy the pitch contour of the spoken sentence. In [6] they model the pragmatics of the speech to classify SLI, ASD and a control group.

In this work pitch, energy, duration, F1-F4 formants, shifted-delta cepstral coefficients (SDC), amplitude modulation (AM) modulation index, and speaking rate are modelled to extract the prosody and long-term information of the speech signal. For pitch, energy, and formants contours over syllable-like intervals are modelled with Legendre polynomials, in a similar approach to [7]. Then, the information from the polynomial coefficients and from the rest of features is extracted in two different ways. First the mean, standard deviation, maximum, and minimum are calculated and used directly in the classifier. Second, the representations in a factor analysis (FA) subspace, or iVectors [8], are obtained and used as features.

The approach is tested on the *Interspeech 2013 Computational Paralinguistics Challenge (ComParE), Autism Subchallenge* [9]. This is the fifth evaluation organized at the same time as Interspeech by a consortium of Universities [1] to share knowledge among speech researchers about different speech tasks like emotion, social signals, or autism, among others. In the case of the autism subchallenge, the organizers provide a baseline [9] over the Child Pathological Speech Database (CPSD) [5]. Two tasks are defined in the subchallenge: *typicality* and *diagnosis*. In the first two classes have to be classified, typically developing (TD) children or control group, and atypical developing (ATY) children; in the second four classes have to be classified, pervasive developmental disorders (PDD) mainly including AD, PDD-NOS, SLI or DYS, and the control group or TYP. For the baseline the auhors use a bench of statistical measures obtained over a set of features including energy, spectral, cepstral, voicing related, harmonics-to-noise ratio, spectral harmonicity, and psychoacoustic sharpness. 6373 features in total obtained with openSmile[2]. As classifier a support vector machine (SVM) is used. Our work compares the performance of the features presented here alone, and combined with the baseline features, using the same SVM classifier as the baseline system. Thus the

[1]http://emotion-research.net/
[2]http://opensmile.sourceforge.net/

Speech Parameter	Stress	Rhythm	Intonation	Articulation
Acoustic Feature	Energy Speaking Rate Modulation Index Duration SDC	Speaking Rate Duration SDC	Pitch SDC	Formants SDC

Table 1: Relation between acoustic parameters and speech parameters

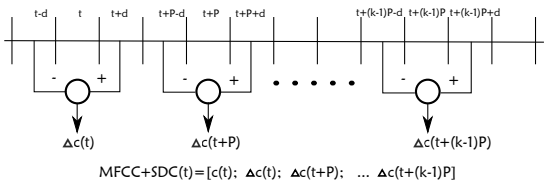

Figure 1: SDC features stacked with MFCC at frame *t* for parameters *N-d-P-k*

differences in perfomance are only due to the different information extracted from the signal.

In the rest of the paper a description of the new features can be found in Section 2, the different modelling of the features is explained in Section 3, including the statistical and the iVector modellings, the results are reflected in Section 4, and the conclusions are drawn in Section 5.

2. Feature Description

The main idea of our features is to capture long-term information of the speech to model the prosody, the tone, the vowel harmony, and in general the speaking style of the four classes. It has been shown that the abnormal use of prosody is a distinguishing property of ASD patients [10, 11], and our hypothesis is that the tone, the vowel harmony, and the speaking style are also affected. In table 1 it is shown with which features each speech parameter is represented. In the following sections these features are described.

2.1. Prosodic Features

Prosody is encoded in syllable length, loudness, and pitch, attributes that make humans perceive rhythm, stress, and intonation. Therefore, our features include pitch to model intonation, energy to model stress, and the number of voiced frames in the current segment to model rythm and stress. The pitch and energy are extracted using the Snack Sound Toolkit [3], and they are obtained in 7.5 ms length windows every 10 ms.

2.2. Formant Modelling

Formants are resonance frequencies of the vocal tract and depend heavily on the position adopted by the speech articulators. The central frequencies of F1, F2, F3, and F4 are used for our experiments. They are extracted with a robust formant tracking algorithm previously proposed in [12]. This method makes use of the roots of the polynomial of a linear predictive coding (LPC) as formant candidates and of a beam-search algorithm for selecting the best combination. The selection is based on a defined cost function, which makes use of information about local and neighbor frames using trajectory functions. The formants are obtained in 49 ms length windows every 10 ms.

2.3. Legendre Polynomials

To model the long-term information of the prosodic features and formants, Legendre polynomials are used to create a regression curve over their contours along a given interval. A decision that has to be made is length of these intervals. Given that prosody is a suprasegmental aspect of speech, being encoded in several speech frames, the contour being modelled should include about tens or hundreds of milliseconds. Fixed segments of 200 ms are created to approximate the average syllable length [13]. For formants the same length is used which was also successfully applied in [7]. The order of the polynomial is 5, thus 6 coefficients are extracted for the pitch, 6 for the energy, and 6 for every formant. Only voiced segments are used to calculate the coefficients, and the number of those within a given interval are concatenated to the coefficients to build the final feature vector representing that interval. Then these coefficients replace the pitch, energy, and formants. The Legendre polynomial approximation is given by

$$f(t) = \sum_{i=0}^{M} a_i P_i(t) \tag{1}$$

where $f(t)$ is the contour being modeled and $P_i(t)$ is the *i*th Legendre polynomial. Each coefficient a_i represents a characteristic of the contour shape: a_0 corresponds to the mean, a_1 to the slope, a_2 to the curvature, and higher order represents more precise detail of the contour.

2.4. Shifted-Delta Cepstral Coefficients

The SDC features are created by stacking delta cepstra computed across multiple speech frames [14]. This gives information about future frames, i.e., long-term information is considered. Four parameters define the SDC: *N*, *d*, *P*, and *k*. *N* is the number of cepstral coefficients computed at each frame, *d* represents the time shift for the delta computation, *P* is the shift between consecutive blocks, and *k* is the number of blocks whose delta coefficients are concatenated. We use a 7-1-3-7 configuration, and stack also the mel-frequency cepstral coefficients (MFCC), so our final vector has 56 dimensions. A graphical representation of this parametrization can be seen in figure 1.

2.5. AM Modulation Index

The AM modulation index (MI) is a measure of the amplitude variation of the modulation signal in time domain. Our hypothesis is that autistic speakers have less control of the energy and will obtain a higher variability in this parameter. To calculate it, the signal is first rectified and downsampled at 60 Hz. Then the maximums and minimums are searched and the MI is defined as

$$MI = \frac{max(k) - min(k)}{max(k) + min'(k)} \tag{2}$$

where $max(k)$ is the *k*th maximum, and $min'(k)$ is the mean of *(k-1)*th and *(k+1)*th minimums, $min'(k) = \frac{min(k-1)+min(k+1)}{2}$ [15].

2.6. Speaking Rate

Two measures of speaking rate are extracted during the MI calculation. The first is the time between the maximums of the signal downsampled at 60 Hz, and the second is the time between the minimums. Our hypothesis is that ASD and SLI groups have different speaking rate to the control group.

[3] http://www.speech.kth.se/snack

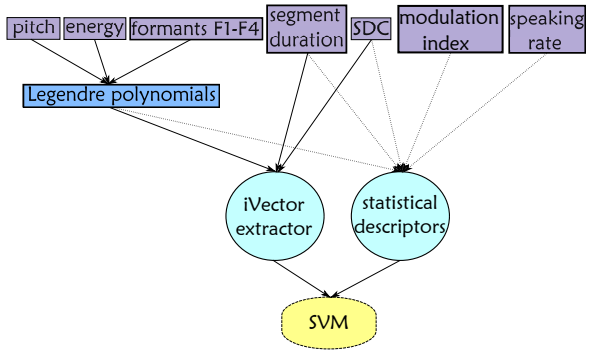

Figure 2: Schematic view of features and their modelling before the SVM

3. Feature Modelling

The classifier used in our experiments is the SVM given in the baseline. As input SVM needs a single vector per utterance. In this section the modellings applied to convert our features into a single vector are detailed. Two different conversions are used, the first is based on a statistical analysis of the features, and the second is based on the iVector approach. In figure 2 a representation of the whole feature structure from the extraction to the compression to be used with SVM is shown.

3.1. Statistical Descriptors

In this approach the mean, standard deviation, maximum, and minimum of the features extracted from each file are computed. It is used with the Legendre polynomials of the pitch, of the energy, and of the formants, with the SDC, and with the MI.

3.2. Factor Analysis Front-End

FA is a modelling technique based on maximum likelihood (ML). Assuming that the data can be modelled by a Gaussian mixture model (GMM) distribution, FA further assumes that the main variability of the signal lies in a low-dimension subspace. Being m_0 a supervector built by concatenating the means of each Gaussian component of the GMM, the means of the FA model for utterance s, $m(s)$, are obtained as

$$m(s) = m_0 + Ti(s), \tag{3}$$

where T is a KxD matrix which translates the vectors from the low-dimension total variability space to the high-dimension space where the model $m(s)$ lies, being D the vector dimension and $K = N$xC the dimension of the supervector, with N the dimension of the input features (Legendre polynomials or SDC in our case), and C the number of components in the GMM; i is the vector in the total variability subspace, also known as *iVector* in the speaker recognition literature, that has an a priori standard normal distribution $\mathcal{N}(0, 1)$. iVectors have been succesfully applied for speaker and language recognition, and since our aim is to compress all the information of the utterance in a single vector, so it can be used with SVM, they seem a priori a good option for that. Our hypothesis is that the main variabilities of the signal modelled by the total variability subspace contain discriminative information for each class in such a way that iVectors are useful for classification. The training of T is done by alternating an ML step with a minimum divergence (MD) step. The iVector is completely characterized by its posterior distribution conditioned to the sufficient statistics of the GMM, and follow a Gaussian distribution. For a complete review of the iVector approach please read [8].

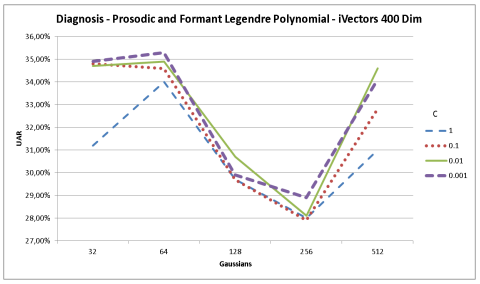

Figure 3: UAR (%) for the diagnosis task. Sweep over number of Gaussians and C for the iVectors extracted from Legendre polynomials calculated from prosodic and formant features

iVector Dimension	50	100	200	400	600
Accuracy	34.9	41.5	42.6	48.2	**48.4**
UAR	26.5	32.1	31.1	**35.3**	**35.3**

Table 2: *Accuracy and UAR (%) in the diagnosis task for the iVector features extracted from Legendre polynomials calculated from prosodic and formant features, for an iVector extractor with 64 Gaussians, and C=0.001*

4. Experiments

The metrics defined to evaluate the system performance are the accuracy and the unweighted average recall (UAR) [9]. In Section 4.1 the SVM is trained on the train dataset and the results are given over the development dataset. This serves us to select the features performing the best, and use them to classify the test dataset, described in Section 4.2.

4.1. Results over Development

We start studying the iVector features for the Legendre polynomials calculated from the prosodic and formant features. For training the GMM and iVector extractor the train and development datasets are used. Hence, when testing the development dataset with the iVector modelling the results are optimistic because of using the development data in the training. The SVM is trained only on the train dataset. The first task is to find the optimal parameters for the number of components in the GMM, the optimal dimension of the iVectors, and the regularizer C of the SVM. For the selection we focus on UAR of the diagnosis task, because it is the metric used to rank systems in the evaluation. Note however that better results could be obtained for the typicality task since the parameters are not optimized for it. In figure 3 it can be observed that for diagnosis the optimal number of Gaussians is 64, and the optimal C is 0.001. In table 2 it is shown a sweep over the iVector dimension. The optimal is 600. For this optimal configuration an UAR of 35.3% is obtained.

The same process is repeated for the iVectors obtained from SDC features. In figure 4 a sweep over the number of Gaussians and over C is shown. The optimal number of Gaussians is 16 and the optimal C is 0.0001. In table 3 it can be checked that the optimal number of iVectors is 400.

In table 4 it can be seen the individual performance of all the proposed features for typicality and diagnosis. Note that for every feature the optimal value of C is different, but for comparison with the baseline, in the typicality task C is set to 0.01, and in the diagnosis task C is set to 0.001. First, it is remarkable that all the proposed features contain useful information, since all of them are above the random result (50% for typicality and 25% for diagnosis). Individually, the best results are achieved for the iVectors obtained from the SDC (ivSDC). As it can be seen, combining these with the statistics extracted from the SDC gives an improvement for typicality but not for diagnosis

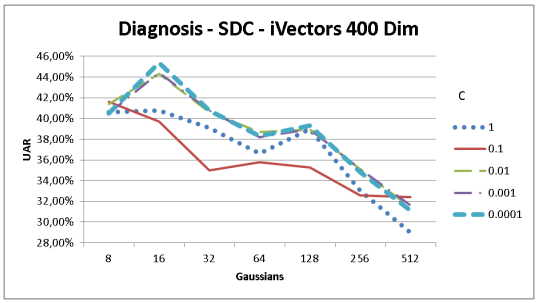

Figure 4: UAR (%) for the diagnosis task. Sweep over number of Gaussians and C for the iVectors extracted from SDC features

iVector Dim	50	100	200	400	600
Accuracy	54.3	56.5	59.5	**61.5**	61.2
UAR	44.7	44.8	44.6	**45.4**	44.6

Table 3: *Accuracy and UAR (%) for the iVector features extracted from Legendre polynomials calculated from prosodic and formant features, for an iVector extractor with 64 Gaussians, and C=0.001*

Task	Typicality		Diagnosis	
Features	Proposed	+Baseline	Proposed	+Baseline
Baseline	-	92.8	-	52.4
statLeg	71.9	93.0	41.0	53.4
ivLeg	67.1	92.6	35.3	52.9
ivLeg+statLeg	73.7	92.4	41.5	53.6
statSDC	78.4	92.1	38.1	52.4
ivSDC	81.9	92.7	44.4	52.7
ivSDC+statSDC	85.0	92.1	43.3	**54.8**
modspkr	58.8	92.9	34.9	52.4
statLeg+statSDC	81.6	92.6	41.3	53.1
statLeg+statSDC+modspkr	81.7	92.4	42.2	53.1
ivLeg+ivSDC+modspkr	80.0	92.9	42.1	53.1
statLeg+statSDC+modspkr+ivSDC	86.5	92.3	**45.1**	53.7
statLeg+statSDC+modspkr+iLeg	83.0	92.6	41.6	54.6
statLeg+statSDC+modspkr+ivSDC+ivLeg	**87.6**	**93.1**	44.9	54.2

Table 4: *UAR (%) for the proposed features alone (column "Proposed") and combined with the baseline (column "+Baseline") in the typicality (C=0.01) and diagnosis (C=0.001) tasks. The abbreviations stand for:*
ivLeg: 600-dimension iVectors of Legendre polynomials from prosodic and formant features (GMM with 64 Gaussians).
statLeg: statistics of Legendre polynomials from prosodic and formant features.
ivSDC: 400 dimension iVectors of SDC (GMM with 16 Gaussians).
statSDC: statistics of SDC.
modspkr: modulation index and speaking rate parameters.

(ivSDC+statSDC). Regarding the prosodic and formant features modelled with Legendre polynomials, the statistical parameters (statLeg) obtain better performance than the iVectors (ivLeg). Their combination improves (statLeg+ivLeg) in both tasks. Adding the modulation index and speaking rate parameters to the combination of the statistical parameters of the prosodic Legendre polynomials and SDC (statLeg+statSDC+modspkr) also gives higher UAR in both tasks than any of the features alone. This is not the case when modelling with iVectors instead of statistical parameters the previous combination (ivLeg+ivSDC+modspkr). Finally, combining all the features together (statLeg+statSDC+modspkr+iSDC+ivLeg) gives the highest performance for the proposed features in the typicality task, while for diagnosis it is better not to include the iVectors of the Legendre polynomials (statLeg+statSDC+modspkr+iSDC) .

In the columns marked with "+Baseline" the proposed features are combined with the baseline features. It is noticeable that the proposed features alone do not reach the results obtained by the baseline features. However, 6373 features are included in the baseline, whereas 1380 are obtained when combining all of the proposed ones. For typicality, combining the baseline with the statistcal paremeters of the prosodic and formant Legendre polynomials already obtains an UAR of 93.0%, and 53.4% for diagnosis. The combination that performs the best for typicality is the one considering all proposed features (statLeg+statSDC+modspkr+ivSDC+ivLeg) with an UAR of 93.1%, a 0.32% improvement over the baseline, whereas for diagnosis it is the combination of iVectors and statistical parameters of SDC (ivSDC+statSDC), that achieves a 54.8%, a 4.58% improvement over the baseline. It is important to say that the given baseline already offers good results in the typicality task and the proposed fetures hardly add new information to it. However for diagnosis, the gains are consistent and we consider this is an interesting starting point for further research.

4.2. Results over Test

After the analysis of results on the development dataset in table 4, two combinations of features are selected for the evaluation of the test dataset. First, the combination of the baseline and all the proposed features (statLeg+statSDC+modspkr+ivSDC+ivLeg), because it gives the best results on the typicality task. Second, the combination of the baseline and SDC modelled with iVectors and with statistical measurements (ivSDC+statSDC), because it gives the best UAR on the diagnosis task. For system evaluation, only the diagnosis task is taken into account, and the typicality results are solved by mapping from the diagnosis task. In this case, for training the SVM the training and development dataset are used. The obtained UAR can be observed in table 5. As it can be checked, there are no improvemets with our proposed features. Further research is needed to understand why the results between development and test differ.

Features	Typicality	Diagnosis
Baseline	**92.60**	**67.10**
Baseline+ivSDC+statSDC	**91**.60	66.06
Baseline+statLeg+statSDC+modspkr+ivSDC+ivLeg	91.06	64.59

Table 5: *UAR (%) results on the test dataset, C=0.001*

5. Conclusions

This work describes the system presented in the ComParE Autism Subchallenge of Interspeech 2013 by I3A. Suprasegmental information of speech is modelled in different ways to classify ASD and SLI. Prosody, represented by pitch, energy, and duration, is modelled by contours given by Legendre polynomials, to extract information about rythm, stress, and intonation. The modulation index gives information of loudness variation along the speech. The speaking rate is included to further model rythm. And the SDC are a modification of the MFCC to include long-term information of the speech in the current frame. For each utterance, statistical parameters of the previous features, and representations in the total variability space, or iVectors, of the prosodic Legendre polynomials, and of the SDC, are extracted to be used with an SVM classifier. The proposed features are shown to be informative alone, and the combination with a set of baseline features given by the organizers obtains promising improvements over the SVM trained on the baseline features alone on the development dataset.

6. Acknowledgements

This work has been funded by the Spanish Government and the European Union (FEDER) under projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

7. References

[1] L. Wing and J. Gould, "Severe Impairments of Social Interaction and Associated Abnormalities in Children. Epidemiology and Classification," *Journal of Autism and Developmental Disorders*, vol. 9, pp. 11–29, 1979.

[2] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR)*. Washington, DC: American Psychiatric Publishing, 2000.

[3] M. Elsabbagh, G. Divan, Y.-J. Koh, Y. S. Kim, S. Kauchali, C. Marcín, C. Montiel-Nava, V. Patel, C. S. Paula, C. Wang, M. T. Yasamy, and E. Fombonne, "Global Prevalence of Autism and Other Pervasive Developmental Disorders," *Autism Research: Official Journal of the International Society for Autism Research*, vol. 5, no. 3, pp. 160–79, Jun. 2012.

[4] D. V. Bishop, "Autism and Specific Language Impairment: Categorical Distinction or Continuum?" *Novartis Foundation Symposium*, vol. 251, pp. 213–26; discussion 226–34, 281–97, Jan. 2003.

[5] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, Jean Xavier, D. Cohen, and M. Plaza, "Automatic Intonation Recognition for the Prosodic Assessment of Language-Impaired Children," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1328–1342, 2011.

[6] E. Prud'hommeaux and M. Rouhizadeh, "Automatic Detection of Pragmatic Deficits in Children with Autism," in *3rd Workshop on Child, Computer and Interaction (WOCCI 2012)*, 2012.

[7] D. Martínez, E. Lleida, A. Ortega, and A. Miguel, "Prosodic Features and Formant Modeling for an iVector-Based Language Recognition System," in *ICASSP*, Vancouver, Canada, 2013.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[9] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge : Social Signals , Conflict , Emotion , Autism," in *Interspeech*, Lyon, France, 2013.

[10] R. Paul, A. Augustyn, A. Klin, and F. R. Volkmar, "Perception and Production of Prosody by Speakers with Autism Spectrum Disorders," *Journal of Autism and Developmental Disorders*, vol. 35, no. 2, pp. 205–220, Apr. 2005.

[11] J. McCann and S. Peppé, "Prosody in Autism Spectrum Disorders: a Critical Review." *International journal of language & communication disorders / Royal College of Speech & Language Therapists*, vol. 38, no. 4, pp. 325–50, 2003.

[12] D. Ribas, J. Garcia, A. Miguel, A. Ortega, E. Lleida, and J. Calvo, "Evaluation of a New Beam-Search Formant Tracking Algorithm in Noisy Environments," in *VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, Iberspeech*, Madrid, Spain, 2012.

[13] S. Greenberg, "Speaking in Shorthand A Syllable-Centric Perspective for Understanding Pronunciation Variation," *Speech Communication*, vol. 29, no. 2-4, pp. 159–176, Nov. 1999.

[14] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features," in *ICSLP*, 2002, pp. 89–92.

[15] J. Villalba and E. Lleida, "Detecting Replay Attacks from Far-Field Recordings on Speaker Verification Systems," in *COST 2101 European Workshop, BioID 2011*. Brandenburg: Springer Berlin / Heidelberg, 2011, pp. 274–285.