Interspeech 2018
2-6 September 2018, Hyderabad

Self-similarity matrix based intelligibility assessment of cleft lip and palate speech

Sishir Kalita[1], *S. R. M. Prasanna*[1,2], *S. Dandapat*[1]

[1]Indian Institute of Technology Guwahati, Guwahati-781039, India
[2]Indian Institute of Technology Dharwad, Dharwad-580011, India

(sishir, prasanna, samaren)@iitg.ernet.in

Abstract

This work presents a comparison based framework by exploiting the self-similarity matrices matching technique to estimate the speech intelligibility of cleft lip and palate (CLP) children. Self-similarity matrix (SSM) of a feature sequence is a square matrix, which encodes the acoustic-phonetic composition of the underlying speech signal. Deviations in the acoustic characteristics of underlying sound units due to the degradation of intelligibility will deviate the CLP speech's SSM structure from that of normal. This degree of deviations in CLP speech's SSM from the corresponding normal speech's SSM may provide information about the severity profile of speech intelligibility. The degree of deviations is quantified using structural similarity (SSIM) index based measure, which is considered as the representative of objective intelligibility score. The proposed method is evaluated using two parameterizations of speech signals: Mel-frequency cepstral coefficients and Gaussian posteriorgrams and compared with dynamic time warping (DTW) based intelligibility assessment method. The propsoed SSM based method shows better correlation with the perceptual ratings of intelligibility when compared to DTW based method.

Index Terms: Cleft lip and palate speech, intelligibility, self-similarity matrix, dynamic time warping.

1. Introduction

Cleft lip and palate (CLP) is one of the most common congenital disorders of the oro-facial region. In this case, individuals with CLP exhibits several speech-related disorders such as articulation errors, hypernasality, and nasal air emission, which lead to poor speech intelligibility [1–3]. Primary articulation errors which affect the intelligibility are maladaptive compensatory articulation, nasalized consonants, weak pressure consonants [4, 5]. The deviations in the articulatory pattern of sound units lead to the significant deviations in acoustic-phonetic cues of the underlying sentence.

In the clinical environment, speech-language pathologist's assess the speech intelligibility using perceptual evaluation based subjective methods [6–8]. The subjective quantification of intelligibility is generally carried out by evaluating the number of correctly uttered words/sentences among the total number of words/sentences used for the evaluation and/or overall articulation capability of the CLP individual [5, 6, 9]. Perceptual evaluation is considered as the gold standard for intelligibility assessment in the clinical setting; though, it has several inherent shortcomings, e.g. biased judgment, need trained experts, and time-consuming process [5, 10]. Thus, an automatic method to objectively quantify the speech intelligibility is always a requirement in this direction to assist the SLPs during the diagnosis and therapy process.

Currently, researchers have explored automatic speech recognition (ASR) based intelligibility measure for German and Italian languages [4, 5]. In ASR based systems, word error rate (WER) is considered to quantify the speech intelligibility and found a high correlation with respect to SLPs perceptual scores. Although ASR based systems provide high correlation, a large amount of annotated data is needed to build acoustic models and language models. Super-vectors based support vector regression is also explored to quantify the CLP speech intelligibility, without utilizing the transcribed speech data [11]. The comparison based approaches using dynamic time warping (DTW) and self-similarity matrices (SSMs) are also explored in other speech applications, where annotated dataset and supervised models are difficult to obtain [12–15].

In ASR based methods, acoustic models are built using normative adult data using MFCC features and adapted for children speech. The acoustic mismatch itself is a great challenge in children ASR, and hence, intelligibility assessment using these systems may not be very reliable for SLPs. Moreover, the generalization of these systems while porting from one language to another for speech assessment may be difficult in case of low resource scenario. In clinical settings, SLPs use specially designed speech stimuli (sentences and words) to assess the CLP speech intelligibility. Thus, a comparison based framework which utilizes the knowledge about the acoustic-phonetic composition of underlying speech stimuli may be helpful in this regard. The attractiveness of comparison based approaches is that they do not make any assumption about the underlying linguistic information [16].

Motivated from the prior discussion, a framework based on self-similarity matrices (SSMs) comparison between normal and CLP speech for intelligibility assessment is proposed. SSM based speech sequence comparison has been found very effective in the word discovery task [15, 17]. SSM's spatial structure totally depends on the underlying sequence of acoustic segments (sound units) present in a particular sentence [15]. Thus, in the degraded intelligibility, where acoustic-phonetic characteristics of sound units are deviated, may lead to changes in the SSMs structure that of normal speech. The deviations of CLP speech's SSMs are expected to correlate with the loss of speech intelligibility. Structural similarity (SSIM) index [18] based image comparison approach is applied to quantify the deviations in the SSMs in the CLP speech. The proposed SSM based measure is compared with DTW accumulated distance based intelligibility measure. Two different features are explored in this works to evaluate the effectiveness of the proposed system: Mel-frequency cepstral coefficients (MFCCs) [19] and Gaussian posteriorgrams [20].

The remainder of this paper is organized as follows: Section 2 provides a brief description of the dataset used in this work and sentence level perceptual evaluation of intelligibility.

10.21437/Interspeech.2018-1125

Table 1: *Description of CLP and normal speakers*

	CLP	Normal
# Total	41	40
# Female, # Male	16, 25	20, 20
Age ($\mu \pm \rho$)	8.79 ± 1.94	9.8 ± 1.42

Table 2: *Correlations of the individual SLPs (raters) to the mean of the other SLPs*

SLPs (Rater)	Mean to other raters		
	ρ	p value	κ
Rater DP	0.80	$p < 0.001$	0.61
Rater NI	0.79	$p < 0.001$	0.60
Rater GI	0.81	$p < 0.001$	0.63

Section 3 discussed the detailed methodology of the proposed system. Experimental results and discussion about the performance evaluation of proposed method are presented in section 4. Finally, Section 5 concludes the paper by providing the summary and future scopes of the work.

2. Database and Perceptual Evaluation

Speech samples of both CLP and normal groups used in this current work are collected from All Indian Institute of Speech and Hearing (AIISH), Mysore, India. All the children with the cleft have undergone primary surgery and do not have other congenital disorders and developmental problems. CLP children with adequate language abilities are only considered for the study. Normal children with matched age and gender, having proper speech and language characteristics are served as controls for the study. Description of the speakers is given in Table 1. Before the recording, ethical consent is obtained from the parents of each group of speakers.

10 phonetically balanced sentences with rich in obstruent consonants are used in this work for the performance evaluation of the proposed algorithm, which is shown in Table 3. These sentences are designed by SLPs of AIISH, Mysore for intelligibility assessment of Kannada CLP individuals. Speech samples are recorded in a sound-treated room using a directional microphone (Bruel & Kjaer) with a sampling frequency of 44 kHz and $16-$bit resolution on a mono channel. The microphone is kept at a distance of 15 cm from each child while recording. For each sentence, $2 - 3$ sessions of recording are conducted for both normal and CLP groups. The database consists of around 1000 CLP speech utterances, while around 1100 normal speech utterances. Three SLPs of AIISH, Mysore who are having around 5 years of experiences in the field of CLP speech evaluation, assess the sentence level intelligibility by perceptual evaluation method. SLPs provide sentence level intelligibility score for each sentence in the scale ranged from 0 to 3, where, $'0'$ = near to normal, $'1'$ = mild, $'2'$ = moderate, $'3'$ = severe. Higher rating value indicates loss of intelligibility, while lower value $'0'$ indicates significantly better intelligibility which is close to normal. To compare the agreement of rating Spearman rank correlation coefficient (ρ) and Cohen's kappa (κ) are computed between the score of individual raters to mean of the respective other two raters. From the Table 2, it can be seen that the intelligibility rating is quite reliable to consider as the ground truth ($p < 0.01$). The median value of the three rater's scores is considered as the ground truth for the current work.

Table 3: *Description of sentence level stimuli used for intelligibility assessment (Written in IPA).*

S1 *kage kalu kapːu*, **S2** *gita bega hogu*, **S3** *dana dari tapːitu*,
S4 *baːlu tabala barisu*, **S5** *beɖa kaɖige odida*, **S6** *sartita kattari taa*,
S7 *ʃivana uru kaːfi* **S8** *ʧaʧa ʧapati kodu*, **S9** *paʈa paʈa bavuʈa*,
S10 *taata tabala taa*

3. Methods

In this section, a detailed discussion of the methodology of SSM based intelligibility assessment is presented. The methodology mainly comprises of three main components: the feature extraction, perform SSM based comparison, and intelligibility score computation. A baseline system is also developed using dynamic time warping (DTW) method for intelligibility assessment to compare with the proposed method.

3.1. Preprocessing and feature extraction

In this work, two features are evaluated for the intelligibility assessment, namely (a) MFCCs and (b) Gaussian posteriograms. Initially, all the speech samples are downsampled to 16kHz and pre-emphasized with a factor of 0.97. The energy-based endpoint detection is applied to detect the starting and end points of the sentences. Preemphasized speech signal lies between the detected end-points are short-term processed by hamming window of size 15 msec with a shift of 5 msec. As all the speech samples are downsampled to 16kHz sampling frequency (fs), 40 numbers of Mel filter banks are considered to compute MFCC features. Along with base 13 dimensional MFCCs (excluding $C0$ coefficient), Δ and $\Delta\Delta$ of variants are also augmented, which results 39 dimensional feature vector. The zero mean and unit variance normalized is performed for each feature dimension before the further processing.

3.1.1. Gaussian posteriogram

Gaussian posteriogram (GP) is a vector of posterior probabilities of each component Gaussian for each feature frame. GP based representation provide speaker independent compact statistical representation of the speech signal [16, 20]. To map the extracted feature vectors to GP for each sentence used in this work, sentence specific speaker independent GMMs are build. Since, 10 sentence level stimuli are used (see Section 2), 10 GMMs are trained from unlabeled normal speaker's data. Later, features computed from each sentence are mapped to the GP using the corresponding sentence specific GMM.

3.2. Self-similarity matrix based comparison

The SSM (Φ_X) of a given frame sequence $F = [f_1, f_2, ..., f_n]$ is a square symmetric matrix, which is computed as, $\Phi_X(i, j) = d(f_i, f_j)$, where, d is any dissimilarity metric between two frames f_i and f_j [17]. For MFCC feature euclidean distance based dissimilarity measure is used, while for GP based representations $-log(GP_1.GP_2)$ is used. To get rid of zeros while computing the log, a discounting based smoothing strategy is applied as discussed in Refs. [15, 20]. It is obvious that the diagonal elements of the SSMs are zero, i.e. $\Phi_X(i, i) = 0$. The structure of an SSM of an utterance is completely depended on its underly sequence of acoustic-phonetic units. The structure of SSM gives robust representation of speech against different speech variabilities, such as, noise, and speaker [15, 17]. The consistent similarities of SSMs for sentence $S1$ (see Section 2) across two normal children (female

and male) are shown in the Fig. 1(a) and (b) respectively. A distinct resemblance of shape patterns and local edges of both the SSMs are observed, which are totally depended on the acoustic units composition in the sentence $S1$. Any distortions in the acoustic-phonetic characteristics of the sound units due to deviations in the articulatory precision or maladaptive compensation will lead to change in SSM's structure. In this work, the deformation of SSM in CLP speech is intended to capture by comparing SSM of the normal speech. The information of dissimilarity may reveal the degree of loss of intelligibility in CLP speech. Fig. 1(a-b), (c), (d), (e), and (f) show the SSMs of normal and four CLP speech utterances, i.e. near to normal, mild, moderate, and severe intelligibility levels respectively. In this case, GPs are used to generate the SSMs of the sentence $S1$, where inner product based similarity measure is used to compute the SSM for better visualization. It can be clearly seen from the figure that structure of the SSMs of CLP speech utterance is deviated more as the intelligibility degrades, due to the deviations in the underlying acoustic-phonetic structure of the utterance.

To capture the dissimilarity among reference SSM and test SSM, initially, warping path ($W(P^*)$) between the two frame sequences (F_R and F_T) is computed using DTW method. This $W(p)$ is used to warp F_R and F_T to $F_R^{'}$ and $F_T^{'}$ to obtain SSMs of same sizes. SSIM index based measure is applied to compare the two SSMs, considering them as the grey scale images.

3.3. Dynamic time warping based comparison

Dynamic time warping (DTW) is a method to estimate the optimal match between two feature sequence by using dynamic programing [21]. Let $F_N = (f_{n1}, f_{n1}, ..., f_{nK})$ and $F_C = (f_{c1}, f_{c1}, ..., f_{cM})$ represent the feature sequences of normal and CLP speech, respectively. Where, K and M corresponds to the number frames of normal and CLP speech, respectively. The DTW distance matrix $D_{K\times M}$ is computed using, $D_{K\times M}(i,j) = d(f_{n_i}, f_{c_j})$, where, d corresponds to any dissimilarity measure between normal speech template f_{n_i} and CLP speech f_{c_j}. The best path in the distance matrix ($D_{K\times M}$) is searched starting from $(1,1)$ and ending at (K,M) using the dynamic programming method, which provides minimal accumulated distance. This accumulated distance is considered as the estimated speech intelligibility score.

3.4. Intelligibility score estimation

As discussed in the section 2, 10 different sentence level stimuli are used in this work for the intelligibility assessment. For each sentence level stimuli, we have considered 10 properly articulated reference utterances from the normal speech data. Thus, 10 SSMs for each feature representation comprises the reference templates for each stimuli. Let's consider, $[X_1^s(F_g), X_2^s(F_g), .., X_r^s(F_g), ..., X_{10}^s(F_g)]$, where $1 \leq s \leq 10$, be the reference SSMs of the s^{th} stimuli for the feature F_g ($g \in \{MFCCs, GP\}$). For the n^{th} test SSM of normal or CLP speech $[Y^{j_n}(F_g)]$, corresponding to j^{th} stimuli, where, $j \in \{1, 2, ..., 10\}$, SSIM index based similarity metric with respect to reference SSMs $[X_r^j(F_g)]_{r=1}^{10}$ is computed. Thus, we have 10 dissimilarity values ($[D_r^{j_n}(F_g)]_{r=1}^{10}$) for that test utterance of j^{th} stimuli and mean of 10 dissimilarity values is considered as the estimated intelligibility score (I^{j_n}) of corresponding utterance. Hence,

$$I^{j_n}(F_g) = \frac{1}{10}\sum_{r=1}^{10} D_r^{j_n}(F_g) \quad (1)$$

Table 4: *Spearman rank correlation (ρ) between subjective intelligibility scores and SSM comparision based scores of sentence S1 for different feature representations*

#	Features	SSM based		DTW based	
		ρ	p	ρ	p
1	$MFCCs$	-0.76	< 0.001	0.55	< 0.001
2	GPs	-0.84	< 0.001	0.73	< 0.001

Table 5: *Average of 10 individual sentence level correlations for overall performance evaluation*

#	Features	SSM based		DTW based	
		ρ	p	ρ	p
1	$MFCCs$	-0.74	< 0.001	0.53	< 0.001
2	GPs	-0.82	< 0.001	0.72	< 0.001

Similar procedure is followed for the DTW based method, where 10 DTW distance scores for each test sentence from 10 reference template is averaged to compute the intelligibility score for that sentence.

4. Experimental Results and Discussion

This section describes the experimental results and performance evaluation of the proposed SSM comparison based intelligibility assessment method. Spearman rank correlation (ρ) between SSM based dissimilarity scores and SLPs perceptual rating are computed for the performance evaluation. The correlation between proposed objective intelligibility scores and subjective intelligibility levels are considered initially for the sentence $S1$ and the results of both the methods are shown in the Table 4. From the Table 4 it can be clearly observed that the correlation values are relatively high for SSM based measure than DTW based measure for both MFCCs and GP features respectively, while comparing with perceptual assessment score. Least correlation value is noted in case of MFCCs based DTW. Hence, this measure may not be very reliable, which may be due to the speaker variabilities embedded in MFCC templates. As GPs provide a speaker independent template representation, the correlation is improved in case GP based DTW. However, the SSM is robust against the speaker variability which is very important in the comparison based approaches. Hence, significant improvement is achieved in MFCCs based SSM as compared to MFCCs based DTW. GPs based SSM measure further increases the correlation by adding more robustness against speaker variabilities in a statistical sense, while retaining the phonetic information.

Later, for 10 sentence level stimuli, correlations are analyzed with respect to perceptual scores, and average of 10 individual sentence level correlations are taken for overall system evaluation. Table 5 shows the average correlation for all the sentence level stimuli. Results show higher correlation in case of GPs based SSM method with a correlation coefficient of -0.82 than DTW based method. The advantage of SSM based approaches is that it captures the dissimilarity among mutual parts of the feature sequence which provides a unique pattern in SSMs for underlying acoustic-phonetic composition. Unlike DTW based approach, SSM based comparison method can encode high information variability among compared patterns by capturing the interaction between all parts of the utterances [15, 17].

Though performance is evaluated in terms of computation of correlation between perceptual assessment score and pro-

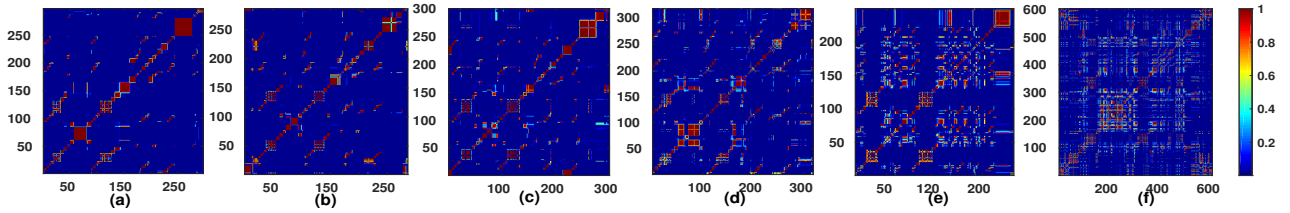

Figure 1: *SSMs structure for normal and CLP speech of sentence S1. (a) normal 1 (female), (b) normal 2 (male), (c) CLP intelligibility level 0, (d) CLP intelligibility level 1, (e) CLP intelligibility level 2, and (f) CLP intelligibility level 3 respectively.*

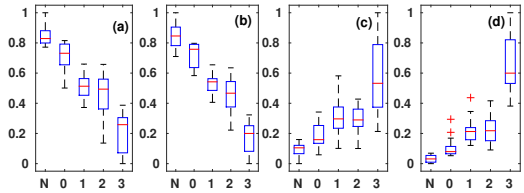

Figure 2: *Box plots of SSM based measure (a, b) and DTW based measure (c, d) for different level of intelligibility in case of sentence* **S1** *for MFCCs and GP based features respectively.*

Table 6: *Results of $MFCCs$ and GPs for both the methods with mean (μ) values for different levels of intelligibility and inter group KL divergence in case of sentence* **S1**

Features	SSM Based (μ)		DTW Based (μ)	
Groups	MFCCs	GPs	MFCCs	GPs
Normal	0.85	0.86	0.092	0.03
0	0.70	0.71	0.18	0.10
1	0.52	0.54	0.31	0.21
2	0.46	0.45	0.29	0.23
3	0.19	0.18	0.58	0.64
	Comparison between the groups (KL divergence)			
Normal vs 0	3.55	2.67	3.05	7.61
0 vs 1	4.90	6.74	2.06	2.5
1 vs 2	0.98	2.04	0.29	0.19
2 vs 3	3.92	6.73	0.96	14.16

posed objective measure, it is important to see the significance of inter-group discrimination (different intelligibility levels and normal) capability. Qualitative discrimination among different groups is shown using box plots for both the SSM based and DTW based methods in Fig. 2. Values of both the intelligibility measures are mapped in between 0 and 1 using min-max normalization prior plotting for all the features. It can be observed from the Fig. 2 that discrimination between groups is more in case of GP based SSM method and least in case of MFCCs based DTW. GP based SSMs method provide some discrimination among the groups 1 and 2, while others almost failed to discriminate them. To quantify the discrimination, mean values of the measures for each group are shown in the Table 6. Further, Kullback-Leibler (KL) divergence among the groups is also studied, apart from the mean analysis (see in Table 6) to quantify. From both mean and KL divergence analysis, it is cleared that for all the features both the SSM and DTW based measures gives significant discrimination among normal Vs 0 and 2 vs 3. However, discrimination among 0 vs 1 and 1 vs 2 is not significant in case of DTW based measure. SSM based measure significantly gives an improvement in terms of discriminating the 0 vs 1 and 1 vs 2. The acoustic-phonetic deviations between 0 vs 1 and 1 vs 2 are very minute, which may not be captured properly by the DTW based measure. Also, the MFCCs capture significant amount of speaker variabilities, which is not able to compensate while matching using DTW. However, SSM provides a robust unique representation of the underlying phonetic content of the sentences and compensate the inherent speech variabilities embedded in the MFCCs. Since, SSM structure is robust against speaker variabilities, it only captures the distortions related to acoustic-phonetic segments due to intelligibility degradation and improves the performance.

It is expected that the proposed method may be helpful in the diagnosis process of CLP individual for the low-resource language context. A proper exploration of patterns in the SSMs may give some insight about the localization of particular sounds misarticulation in the utterance. Though the proposed method shows significant correlation with the perceptual ratings, the complexity arises since separate acoustics models are needed for individual sentences for GP based feature representation.

5. Conclusion and future directions

In this paper, SSM based comparison framework is proposed to objectively estimate the CLP children's speech intelligibility. The primary motivation of the work is to explore an unsupervised way of estimating the speech intelligibility, which does not make any explicit assumptions about the acoustic and linguistic knowledges. MFCCs and GPs based feature representation are explored to evaluate the effectiveness of the proposed method and compared to DTW based method. The estimated objective intelligibility scores are compared with the perceptual rating in terms of correlation analysis. SSM based method gives significantly high correlation while evaluating with respect to perceptual scores, compare to DTW based method. GP based SSMs method provides the highest correlation with significant discrimination among the different intelligibility groups.

In the future work, global intelligibility score of the CLP children is planned to estimated with the help of sentence-level scores. For better modelling of the acoustic units deep belief network (DBN) based postriorgrams can be explored in future.

6. Acknowledgement

Authors would like to thank Prof. M Pushpavathi and expert SLPs of AIISH, Mysore for their valuable contribution in the perceptual evaluation of speech. This work is in part supported by the project grants, for the projects entitled "NASOSPEECH: Development of Diagnostic system for Severity Assessment of the Disordered Speech" funded by the Department of Biotechnology (DBT), Govt. of India and "ARTICULATE +: A system for automated assessment and rehabilitation of persons with articulation disorders" funded by the Ministry of Human Resource Development (MHRD).

7. References

[1] K. Bzoch, "Introduction to the study of communicative disorders in cleft palate and related craniofacial anomalies," *Communicative disorders related to cleft lip and palate. 5th ed. Austin: pro-ed*, pp. 3–66, 2004.

[2] B. J. Costello, R. L. Ruiz, and T. A. Turvey, "Velopharyngeal insufficiency in patients with cleft palate," *Oral and Maxillofacial Surgery Clinics of North America*, vol. 14, no. 4, pp. 539–551, November 2002.

[3] A. Kummer, *Cleft palate & craniofacial anomalies: Effects on speech and resonance*. Nelson Education, 2013.

[4] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.

[5] A. Maier, C. Hacker, E. Noth, E. Nkenke, T. Haderlein, F. Rosanowski, and M. Schuster, "Intelligibility of children with cleft lip and palate: Evaluation by speech recognition techniques," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, 2006, pp. 274–277.

[6] T. L. Whitehill, "Assessing intelligibility in speakers with cleft palate: A critical review of the literature," *The Cleft Palate-Craniofacial Journal*, vol. 39, no. 1, pp. 50–58, 2002, pMID: 11772170.

[7] D. Sell, A. Harding, and P. Grunwell, "A screening assessment of cleft palate speech (great ormond street speech assessment)," *International Journal of Language & Communication Disorders*, vol. 29, no. 1, pp. 1–15, 1994.

[8] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone, and T. L. Whitehill, "Universal parameters for reporting speech outcomes in individuals with cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 45, no. 1, pp. 1–17, 2008.

[9] M. Scipioni, M. Gerosa, D. Giuliani, E. Noth, and A. Maier, "Intelligibility assessment in children with cleft lip and palate in italian and german," in *Interspeech 2009*, 2009.

[10] M. Schuster, A. Maier, T. Bocklet, E. Nkenke, A. Holst, U. Eysholdt, and F. Stelzle, "Automatically evaluated degree of intelligibility of children with different cleft type from preschool and elementary school measured by automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 76, no. 3, pp. 362–369, 2012.

[11] T. Bocklet, A. Maier, K. Riedhammer, and E. Noth, "Towards a language-independent intelligibility assessment of children with cleft lip and palate," in *In Proc. WOCCI 2009*, November 2009, pp. 4366–4369.

[12] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, May 2012, pp. 382–387.

[13] J. C. Vasquez-Correa, J. R. Orozco-Arroyave, and E. Noth, "Word accuracy and dynamic time warping to assess intelligibility deficits in patients with parkinsons disease," in *2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, Aug 2016, pp. 1–5.

[14] R. Ullmann, M. M. Doss, and H. Bourlard, "Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4924–4928.

[15] A. Muscariello, G. Gravier, and F. Bimbot, "Towards robust word discovery by self-similarity matrix comparison," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5640–5643.

[16] Y. Zhang, "Unsupervised speech processing with applications to query-by-example spoken term detection," Ph.D. dissertation, Massachusetts Institute of Technology, 2013.

[17] A. Muscariello, G. Gravier, and F. Bimbot, "Unsupervised motif acquisition in speech via seeded discovery and template matching combination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2031–2044, Sept 2012.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[19] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[20] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, Nov 2009, pp. 398–403.

[21] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, ser. Prentice-Hall Signal Processing Series: Advanced monographs. PTR Prentice Hall, 1993.