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

Sishir Kalita¹, S. R. M. Prasanna¹², S. Dandapat¹

¹Indian Institute of Technology Guwahati, Guwahati-781039, India
²Indian 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 proposed 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-vecors-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 work 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.
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 3 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, 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, 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.

Table 1: Description of CLP and normal speakers

<table>
<thead>
<tr>
<th></th>
<th>CLP</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td># Total</td>
<td>41</td>
<td>40</td>
</tr>
<tr>
<td># Female, # Male</td>
<td>16, 25</td>
<td>20, 20</td>
</tr>
<tr>
<td>Age (µ ± ρ)</td>
<td>8.79 ± 1.94</td>
<td>9.8 ± 1.42</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Rater</th>
<th>Value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP</td>
<td>0.80</td>
<td>p &lt; 0.001 0.61</td>
</tr>
<tr>
<td>NI</td>
<td>0.79</td>
<td>p &lt; 0.001 0.60</td>
</tr>
<tr>
<td>GI</td>
<td>0.81</td>
<td>p &lt; 0.001 0.63</td>
</tr>
</tbody>
</table>

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. Methods

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 end-point 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 (f_s), 40 numbers of Mel filter banks are considered to compute MFCC features. Along with base 13 dimensional MFCCs (excluding C0 coefficient), Δ and ΔΔ 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 posterior (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^n}) of a given frame sequence F = [f_1, f_2, ..., f_n] is a square symmetric matrix, which is computed as, Φ_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_i, GP_j) 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. Φ_X(i, i) = 0. The structure of an SSM of an utterance is completely depended on its underlying 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 \((F_R, F_T)\) 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 \(F_N = (f_{n1}, f_{n2}, \ldots, f_{nK})\) and \(F_C = (f_{c1}, f_{c2}, \ldots, f_{cM})\) represent the feature sequences of normal and CLP speech, respectively. Where, \(K\) and \(M\) correspond to the number frames of normal and CLP speech, respectively. The DTW distance matrix \(D_{K \times M}\) is computed using,

\[
D_{K \times M}(i, j) = d(f_{ni}, f_{cj}), \text{where, } d \text{ corresponds to any dissimilarity measure between normal speech template } f_{ni} \text{ and CLP speech } f_{cj}. \]

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, \(X_{i}^j(F_g), X_{i}^j(F_p), \ldots, X_{i}^j(F_g), \ldots, X_{i}^j(F_p))\), where \(1 \leq s \leq 10\), be the reference SSMs of the \(s^{th}\) stimuli for the feature \(g\) \((g \in \{MFCCs, GP\})\). For the \(s^{th}\) test SSM of normal or CLP speech \(Y_{i}^j(F_g)\), corresponding to \(j^{th}\) stimuli, where, \(j \in \{1, 2, \ldots, 10\}\). SSIM index based similarity metric with respect to reference SSMs \(X_{s}^j(F_g)\) is computed. Thus, we have 10 dissimilarity values \((D_{s}^{i,j}(F_g))_{r=1}^{10}\) for that test utterance of \(j^{th}\) stimuli and mean of 10 dissimilarity values is considered as the estimated intelligibility score \((I_{s}^{i,j})\) of corresponding utterance. Hence,

\[
I_{s}^{i,j} = \frac{1}{10} \sum_{r=1}^{10} D_{s}^{i,j}(F_g) \quad (1)
\]

### 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 \((\rho)\) 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 \([15, 17]\).

Table 4: Spearman rank correlation \((\rho)\) between subjective intelligibility scores and SSM comparison based scores of sentence S1 for different feature representations

<table>
<thead>
<tr>
<th># Features</th>
<th>SSM based</th>
<th>DTW based</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCCs</td>
<td>(-0.76)</td>
<td>(&lt; 0.001)</td>
</tr>
<tr>
<td>GPs</td>
<td>(-0.84)</td>
<td>(&lt; 0.001)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th># Features</th>
<th>SSM based</th>
<th>DTW based</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCCs</td>
<td>(-0.74)</td>
<td>(&lt; 0.001)</td>
</tr>
<tr>
<td>GPs</td>
<td>(-0.82)</td>
<td>(&lt; 0.001)</td>
</tr>
</tbody>
</table>

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.
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 acoustic 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 GP 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 posteriorgrams 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


