SPEECH SOUND DISORDER CLASSIFICATION BASED ON TIME-ALIGNED DISSIMILARITY PROFILES

Guilherme Jun Yoshimura **Marcelo Queiroz**
Computer Science Department, University of São Paulo
{jun,mqz}@ime.usp.br

Haydée Fiszbein Wertzner **Danira Francisco**
School of Medicine, University of São Paulo
hfwertzn@usp.br, daniratavares@gmail.com

ABSTRACT

Speech sound disorders (SSD) are characterized by a person's difficulty (or inability) in producing specific sounds or pronouncing certain words correctly. In this project we are dealing with SSD that appear during the development of speech; these are diagnosed by phonologists using specific protocols and comparing a child's utterance of a specific word with a reference pronunciation. In order to help them to detect and speed up diagnosis we propose a classifier based on dissimilarity profiles built out of DTW-aligned MFCCgrams. Unlike usual classifiers based on statistical audio features, this method preserves the temporal sequence of the audio recordings, which usually have different durations. We compare the proposed method with two other SSD classifiers previously used for the same task, one based on the Earth Mover's Distance, and another that uses a relative DTW embedding (minDTW). We present results showing that the proposed method compares favorably with respect to the competitors on a dataset used for SSD diagnosis in children speaking Brazilian Portuguese.

1. INTRODUCTION

Speech sound disorders (SSD) are usually noticed during the development of speech when children fail to use certain sounds by the expected age, or use them improperly, and in adults as results of either speech disorders not treated in early ages or traumatic brain injuries. Children with SSD may have difficulties in auditory perception, phonological representation and/or production of speech sounds. This difficulty can interfere with an individual's communication abilities and may affect academic and professional performance [1].

One of the challenges for speech and language pathologists is the screening phase, where the professional needs to apply a test and then evaluate each word that the patient produces; usually this evaluation is carried out by several experts, and a majority vote criterion is adopted. This is by far the most time-consuming phase of the diagnosis process. The focus of this paper is to present a method that classifies speech sound recordings according to whether the patient appears to have pronounced the word correctly or not, so that the speech and language pathologists can reduce the time spent in diagnosis and proceed earlier to treatment.

The problem of automatic speech sound disorder classification has been studied for over a decade, using classical machine learning tools and considering specific SSDs such as dysarthria/stuttered speech [2–5] and phoneme replacement [5]. Two approaches to the speech classification problems that are closely related to the present work combine MFCC features respectively with the EMD (Earth Mover's distance [6]) and the MinDTW distance [5]. EMD was originally proposed by Rubner et al. [7] to solve image retrieval problems and was used in [6] for speaker identification. EMD was intuitively defined as a measure of effort in transporting mass (i.e. density) from one probability distribution to another. MinDTW [5] on the other hand is a classifier based on Dynamic Time Warping (DTW) that handles audio recordings with different time durations and considers their temporal sequence through DTW-aligned MFCCgrams. Their accumulated dissimilarity measure is then embedded in a relative distance space for testing pertinence to the reference class. MinDTW has been shown [5] to perform better than both HMM and Vector Quantization with Bag-of-Words on the UA-Speech Database [8], which is one of the largest databases available for disordered speech.

In this work we propose a new classifier based on a representation using dissimilarity profiles built from a query MFCCgram that is time-aligned (using DTW) against MFCCgrams within the reference class (i.e. the recordings of typical utterances with no diagnosed SSD). The motivation is to provide the classifier with an interpretable temporal representation that allows phonologists to visualize how utterances evolve in time with respect to their adherence to typical utterances of the same word, which is especially useful in the case of phoneme replacements. We compare the results of this classifier with the two previously mentioned classifiers based on the EMD and the MinDTW distances.

The structure of the paper is as follows. Section 2 presents the database used in the experimental part of this paper. Section 3 formalizes the proposed method and strategy for automatic SSD classification. In section 4 we present and discuss the results of the experiment. Conclusions and future work are presented in Section 5.

2. DATABASE

The database used in the experimental section was provided by the Department of Physical Therapy, Speech, Language and Hearing Sciences, and Occupational Therapy, of the School of Medicine at the University of São Paulo. It was created in order to allow studies of speech sound disorders in Brazilian Portuguese in adults and children [9].

The subset of the database used in the current experiment was manually prepared (i.e. segmented and labeled) in order to allow batch processing for automatic training and classification. It is composed of 200 recordings of two Portuguese words widely used in SSD diagnosis in children, the words *sapo* (pronounced [ˈsapu], meaning *frog*) and *chave* (pronounced [ˈʃavi], meaning *key*). These recordings were obtained during the master's study of the fourth author [1] . Participants were 21 children aged 5 thru 11, with no familial or personal history of diagnosed or suspected auditory, otologic or neurological disorder or injuries, and no previous speech-language interventions.

The recordings are tagged according to whether they display any form of SSD, based on their score on the Phonology Proof of Child Language Test – Percentage of Consonants Correct (ABFW–PCC) [10], which is the officially adopted protocol for diagnosing SSD in Brazilian Portuguese. For each of the two words used in the experimental part of this paper, 60% were labeled as reference recordings (typical utterances, no SSD) and 40% as presenting some form of SSD.

We acknowledge that the dataset used in the current experiment is smaller than other databases used for SSD classification, such as the UA-Speech [8] (60 recordings of 765 words by 20 participants). Our main interest is to explore the feasibility of using our representation and classifier in the screening phase of diagnosis, in close collaboration with phonologists that apply the ABFW-PCC protocol in the Brazilian Portuguese language. The judgement, by phonologists, of the adequacy of the dissimilarity profiles in representing specificities of the SSDs considered, requires some level of acquaintance with the recordings and its patients. This has led us to consider the database available at the School of Medicine of the University of São Paulo, whose recordings were unfortunately still lacking segmentation and labeling. This requirement led us to restrict the number of words, while ensuring that the number of recordings for each word was comparable to that of the UA-Speech database (which has 60 recordings per word – we used 100).

3. TADPC AND ITS COMPETITORS

3.1 Time-Aligned Dissimilarity Profile Classifier

The motivation for the method here proposed, called Time-Aligned Dissimilarity Profile Classifier (TADPC), is to produce a unified comparative measure of a given recording with respect to the whole set of reference recordings (i.e. typical, non-SSD utterances of the same word). For each given recording X (which may be marked as with or without SSD), it builds a dissimilarity profile based on DTW time-aligned comparisons of X with each reference recording. These profiles are then summarized into a single time-aligned dissimilarity profile, which serves as basis for training the classifier. Figure 1 represents the steps of the proposed method to produce the Time-Aligned Dissimilarity Profile (TADP) of X, a process which is detailed in the sequel.

In the first step, recordings are represented by MFCC-grams, obtained using segments of 2048 audio samples with 75% overlap and 12 mel-frequency cepstral coefficients. Each reference recording Y is then time-aligned to X using DTW [11], and a temporal profile is created based on the dissimilarity values over the optimal alignment path identified by the Viterbi algorithm within the DTW matrix. These profiles use as time axis the indices referring to the recording X, so that they can all be superimposed within a unified time span corresponding to the duration of X.

The second step corresponds to unifying these profiles into a single Time-Aligned Dissimilarity Profile (TADP). This is done by considering a parameter $\alpha \in [0, 1]$ that represents a percentile for the dissimilarity values within each time frame. Specifically, each dissimilarity profile P_Y obtained from the comparison of X and reference recording Y defines a dissimilarity $P_Y(i)$ on each frame i. Considering for each frame i the set

$$D(i) = \{P_Y(i) \mid \forall \text{ reference recordings } Y\} \qquad (1)$$

we define x_i as the α percentile of $D(i)$. This defines a unified profile $(x_1, \ldots, x_n)$ for X, called TADP, that correspond to a time-varying statistic reflecting all individual dissimilarity profiles P_Y. Figure 2 presents an example of the unified TAPD obtained from the set of profiles in Figure 1 using an 80% percentile. The actual value of this percentile is chosen during the training phase in order to optimize the F-measure of the resulting classifier when applied to the known training data.

Finally, from the resulting TADP $(x_1, \ldots, x_n)$ of X we take the average dissimilarity

$$\text{TADPDistance}(X) = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (2)$$

as a distance-like measure of pertinence of X to the reference class. The rationale is that MFCCgrams of reference recordings are not so different from one another after time-alignment, and so the dissimilarity values tend to be low overall, producing a low TADPDistance. On the other hand, recordings with some form of SSD will produce dissimilarity spikes in phonemes which do not correspond to the typical utterances in reference recordings, and these would increase the TADPDistance.

A classifier is then obtained by choosing the threshold for the TADPDistance that optimizes the F-measure for classification within the training set. Specifically, we perform a binary search within a range $[\mu_l, \mu_h]$ where μ_l is the minimum TADPDistance for recordings displaying speech

[1] Research protocol 276/13 approved by the ethics committee at the School of Medicine of the University of São Paulo. Terms of informed consent were signed by participants' parents, and all data was anonymized.

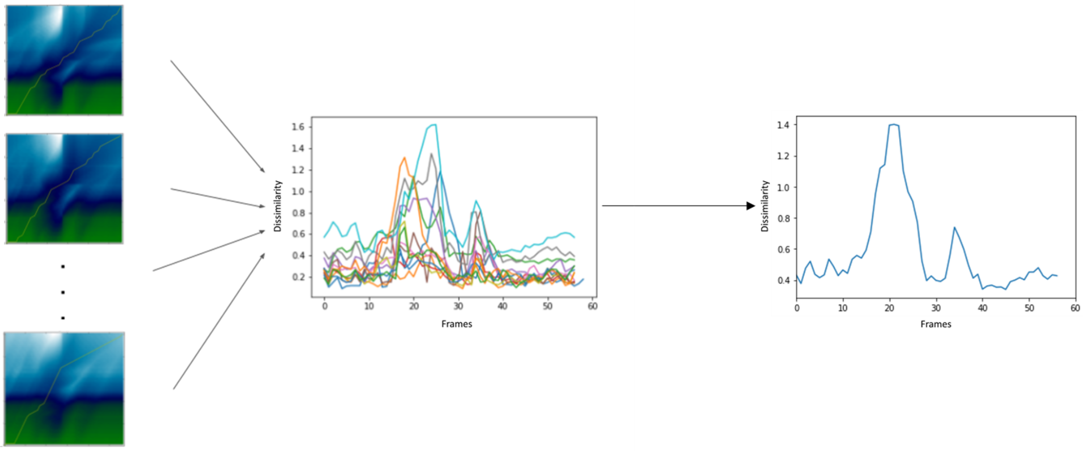

Figure 1. Dissimilarity profiles for a given recording X against all reference recordings.

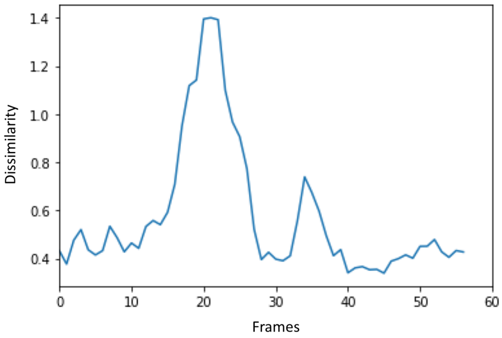

Figure 2. Time-Aligned Dissimilarity Profile of X based on an 80% percentile from the profiles in Figure 1.

sound disorders and μ_h is the maximum TADPDistance for reference recordings. If the TADPDistance values corresponding to both classes do not overlap, in other words if $\mu_l > \mu_h$, then we define the threshold simply as $\frac{\mu_l+\mu_h}{2}$. This resulting method is called Time-Aligned Dissimilarity Profile Classifier (TADPC).

3.2 Classifier based on the Earth Mover's Distance

The classifier based on the Earth Mover's Distance (EMD) is built from the same dissimilarity values appearing in the optimal Viterbi path within the DTW matrix, but it compares the probability density functions (pdfs) of these dissimilarity values. The probability density functions remove the temporal aspects of the dissimilarity profiles while still allowing the distinctions of profiles with unusually high dissimilarity values, as would correspond to recordings of disturbed speech, in comparison to profiles obtained from other non-disturbed recordings.

In order to establish a pdf corresponding to the reference (non-SSD) recordings, we align each pair of reference recordings using DTW, and collect the dissimilarity values appearing in the optimal Viterbi path within the DTW matrix. These dissimilarity values are expected to be lower than the ones appearing in comparisons of atypical (SSD) utterances and typical (non-SSD) recordings, so the resulting pdf will have most of its mass concentrated in the low region.

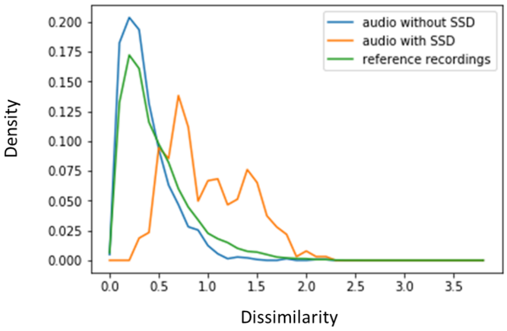

Figure 3. Probability density functions for dissimilarity values of reference recordings (green), a recording without SSD (blue), and a recording with SSD (orange), used by the EMD classifier.

For each given recording X, a pdf of dissimilarity values is obtained in a similar way: X is compared to all reference recordings using DTW, and the values appearing in the optimal Viterbi path within the DTW matrix define the pdf corresponding to X. This pdf is expected to be closer (with respect to EMD) to the pdf of the reference class if X is a typical utterance, and somewhat different when X is atypical (i.e. with SSD). Figure 3 shows an example of three pdfs, one for the whole set of reference recordings (green), another for a given recording without any SSDs (blue) and a third one for a given recording with some form of SSD (orange).

A pertinence value for a given recording X is then defined by the EMD between X and the pdf of the reference recordings. Based on these EMD values we may obtain a threshold to separate the classes using a strategy similar to the one defined for the TADPC. During training, an optimized threshold is sought after so that the classifier achieves the highest F-measure within the training set.

3.3 Classifier based on MinDTW

MinDTW [5] defines a distance from any given recording X to the class of all reference recordings, by considering the reference recording Y closest to X with respect to their DTW distances, as illustrated in Figure 4. Applying this definition to the classes of typical and atypical utterances, a classifier may be obtained by choosing an optimal threshold for the corresponding MinDTW values as was done for the EMD and TADP classifiers. In [5] this method was compared to two other well-known methods, HMM [12] and Vector Quantization with Bag-of-Words (VQ+BoW) [13], and MinDTW achieved an F-measure of 95% on the UA-Speech database [8], against 83% for HMM and 81% for VQ+BoW.

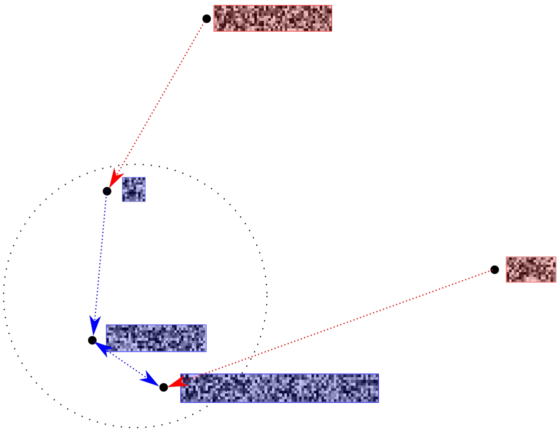

Figure 4. Binary classification using relative DTW embedding (MinDTW). Reproduced from [5].

Compared to TADPC, MinDTW uses only the accumulated dissimilarity expressed by the DTW distance, and establishes the pertinence of a given recording X by comparing X to a single reference recording Y (the one with minimum DTW distance), which acts as a proxy to the class of all reference recordings. TADPC compares with and generalizes MinDTW by combining all dissimilarity curves obtained from the DTW-alignment of X to each reference recording, which are time-aligned and unified as described in Section 3.1. Rather than simply replacing the "Min" in MinDTW by some other statistic (mean/maximum/median), the representation TADP aim to attain as much a global perspective as possible on the relationship between the query and the reference class, by preserving both statistical insertion of dissimilarity values (via percentiles) as well as the temporal evolution of this insertion.

4. EXPERIMENTAL METHODOLOGY AND RESULTS

In this section we discuss the experimental methodology employed to compare the results of TADPC against the EMD and MinDTW classifiers described in Section 3.2. To evaluate the classifiers a K-fold cross-validation was used (with $K = 5$) in order to produce F-measure values. Two summarization strategies were used: a global F-measure [14] calculated from the accumulated amounts of true positives (TP), false positives (FP), and false negatives (FN) over all folds $k = 1, \dots, K$ according to

$$F_{\text{global}} = \frac{2 \cdot \sum_k TP[k]}{2 \cdot \sum_k TP[k] + \sum_k FP[k] + \sum_k FN[k]} \tag{3}$$

and also the mean and standard deviation of the F-measures of individual folds. The global F-measure, denoted here by F-Global, is considered to minimize the bias in comparison with other F-measure summarization strategies [14]. The mean/std format, denoted here by F-Normal, represent a Gaussian model of the values K individual F-measures, allowing statistical tests to be performed in order to compare the performance of the methods.

Table 1 presents the global F-measure and mean/std of F-measures of individual folds for each classifier and each word. It is noticeable that the classifier based on the Earth Mover's Distance obtained the lowest F-measure for both words. One possible reason for the EMD classifier's worse performance is the fact that the probability density functions of dissimilarity values disregard the temporal sequence of the recordings.

In order to compare the performances of the remaining methods, we applied a paired (repeated samples) t-test to the individual F-measures obtained in the K folds of the cross-validation using SciPy `ttest_rel` function. Based on these tests, the only statistically significant differences are that both TADPC and HausdorffDTW performed better than MinDTW for the word "chave" ($p \ll 0.01$).

One plausible explanation that would explain TADPC outperforming MinDTW for the word "chave" relates to specific forms of SSD utterances that are found in these recordings, which are easier to distinguish with respect to most reference recordings, even when there are a few (outlier) reference recordings displaying overall lower dissimilarity values with respect to this particular atypical (SSD) utterance. See for instance Figure 5, where most profiles display high dissimilarity values between frames 5 and 18, but the violet and green profiles remain entirely in the low region. In these cases, MinDTW chooses as score the average dissimilarity of the (outlier) lower profile, mistaking it for a typical reference utterance, whereas TADPC considers the worst profiles according to the percentile parameter α. It should be noted that a classifier that only considered the worst profile had already been proposed in [5], the HausdorffDTW classifier, but its performance was much worse than MinDTW in the UA-Speech dataset.

	sapo ([ˈsapu])		chave ([ˈʃavi])	
	F-GLOBAL (%)	F-NORMAL(%)	F-GLOBAL (%)	F-NORMAL(%)
MinDTW	81.42	81.34 ± 0.07	75.18	75.16 ± 0.03
HausdorffDTW	80.7	80.62 ± 0.06	76.06	76.08 ± 0.02
EMD	76.78	76.33 ± 0.1	64.66	59.23 ± 0.25
TADPC	81.03	80.78 ± 0.05	76.26	76.13 ± 0.03

Table 1. Global F-measure and mean/std of F-measures of individual folds, for each classification method and word.

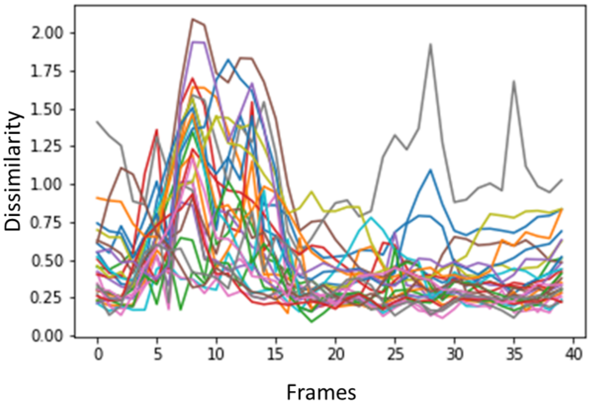

Figure 5. All dissimilarity profiles built from an atypical (SSD) utterance of the word “chave”, with respect to all typical reference recordings of the same word.

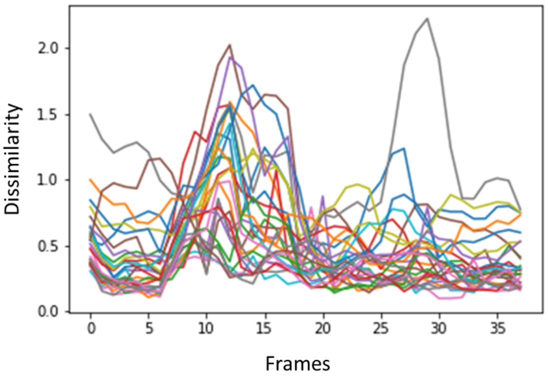

Figure 6. Same as Figure 5 but from another SSD patient.

On the other hand, upon a closer look of the dissimilarity profiles produced by the atypical recordings of the word “sapo”, the high dissimilarity values tend to concentrate in the same intervals in all profiles (see e.g. Figure 7), and so choosing a single curve (as MinDTW does) to compute the score, or alternatively using a percentile-summarized curve as TADPC does, makes less of a difference in the classification results and thus in the final F-measures.

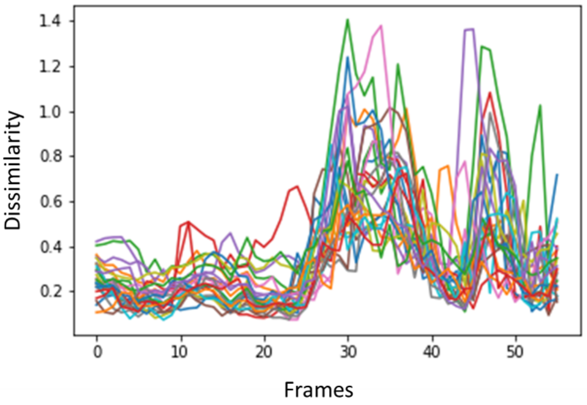

Figure 7. All dissimilarity profiles built from an atypical (SSD) utterances of the word “sapo”, with respect to all typical reference recordings of the same word.

5. CONCLUSION

In this paper we introduced a classifier that uses dissimilarity values of MFCCgrams based on DTW-aligned pairs of recordings. By using the values of dissimilarity along the optimal Viterbi path in the DTW matrices for all reference recordings, we built a classifier that not only preserves the temporal sequence of the dissimilarity profiles (as MinDTW does) but also takes into account variations between utterances within the reference class. TADPC was shown to compare favorably with respect to the EMD-based and the MinDTW classifiers.

The experiment here presented was conducted within a subset of a larger database, as an exploratory step to identify the potential of TADPC to outperform MinDTW in the detection of SSDs in the Brazilian Portuguese language. Extending the experiment to the whole database still depends upon a reasonable amount of manual labor to preprocess several hours of unsegmented recordings into files containing individual words, nor clearly labeled as displaying some type of SSD (for many recordings the SSD tag refers to the patient and not to the specific word, with potential mislabeling).

A well-known issue with DTW-based methods is DTW’s quadratic computational complexity. Nevertheless, its complexity refers to the size of the recordings, which are short utterances of individual words (40 to 50 frames here) in the SSD classification setting, and not to the size of the dataset. The main scalability issue appears when a query has to be DTW-aligned to all reference recordings in the training set. For large datasets, it is advisable to work with a small number of typical recordings for each word in order not allow

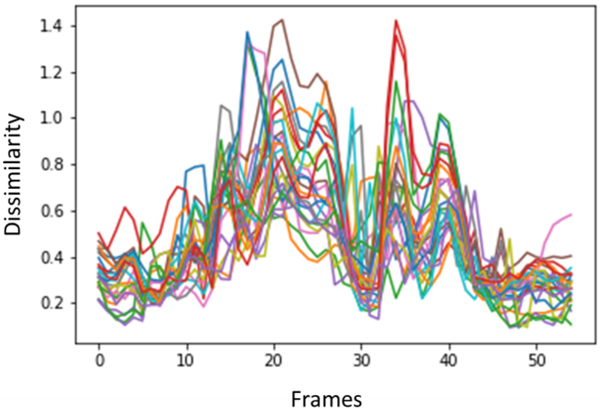

Figure 8. Same as Figure 7 but from another SSD patient.

the fast computation of dissimilarity profiles.

Future work includes tackling challenges that appeared when confronting MinDTW and TADPC for specific disorders, as illustrated in Figures 5 and 7, which shed some light into the variations within the very class of reference recordings, and also exploring the Time-Aligned Dissimilarity Profiles as input for temporal localization (segmentation) of the phonemes displaying variations with respect to the reference pronunciation, in order to help phonology professionals to refine the diagnosis and identify particular subclasses of speech sound disorders.

Acknowledgments

The second author acknowledges having received financial support from CNPq grant 307389/2019-7.

6. REFERENCES

[1] J. McCormack, S. McLeod, L. McAllister, and L. J. Harrison, “A systematic review of the association between childhood speech impairment and participation across the lifespan,” *International Journal of Speech-Language Pathology*, vol. 11, no. 2, pp. 155–170, 2009. [Online]. Available: https://doi.org/10.1080/17549500802676859

[2] K. Ravikumar, R. Rajagopal, and H. Nagaraj, “An approach for objective assessment of stuttered speech using mfcc features,” *ICGST International Journal on Digital Signal Processing, DSP*, vol. 9, pp. 19–24, 01 2009.

[3] M. Wisniewski, W. Kuniszyk-Józkowiak, E. Smolka, and W. Suszynski, “Automatic detection of prolonged fricative phonemes with the hidden markov models approach,” *Journal of Medical Informatics & Technologies*, vol. 11, 01 2007.

[4] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, “Mfcc based recognition of repetitions and prolongations in stuttered speech using k-nn and lda,” in *2009 IEEE Student Conference on Research and Development (SCOReD)*, Nov 2009, pp. 146–149.

[5] M. Queiroz and G. J. Yoshimura, “Relative DTW Embedding for Binary Classification of Audio Data,” in *In: Proceedings of the 15th Sound and Music Computing Conference (SMC 2018)*, Limassol, Cyprus, 2008, pp. 279–286.

[6] S. Kuroiwa, S. Tsuge, M. Kita, and F. Ren, “Speaker identification method using earth mover’s distance for CCC speaker recognition evaluation 2006,” in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, Sep. 2007, pp. 239–254. [Online]. Available: https://www.aclweb.org/anthology/O07-5001

[7] Y. Rubner, L. J. Guibas, and C. Tomasi, “The earth mover’s distance, multi-dimensional scaling, and color-based image retrieval,” in *Proceedings of the ARPA image understanding workshop*, vol. 661, 1997, p. 668.

[8] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, K. W. T. Huang, and S. Frame, “Dysarthric speech database for universal access research,” in *Proceedings of Interspeech*, Brisbane. Australia, 2008, pp. 1741–1744.

[9] D. Francisco and H. Wertzner, “Differences between the production of [s] and [ʃ] in the speech of adults, typically developing children, and children with speech sound disorders: An ultrasound study,” *Clinical linguistics & phonetics*, vol. 31, pp. 1–16, 01 2017.

[10] L. D. Shriberg and J. Kwiatkowski, “Phonological disorders i: A diagnostic classification system,” *Journal of Speech and Hearing Disorders*, vol. 47, no. 3, pp. 226–241, 1982.

[11] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978, pp. 43 – 49 Volume: 26, Issue: 1, Feb 1978.

[12] M. Gales and S. Young, “The application of hidden markov models in speech recognition,” *Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[13] S. Pancoast and M. Akbacak, “Bag-of-audio-words approach for multimedia event classification,” in *Thirteenth Annual Conference of the International Speech Communication Association*, pp. 2105–2108.

[14] G. Forman and M. Scholz, “Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement,” in *SIGKDD Explor. Newsl.*, 2010, pp. 49–57, Volume 12, Issue 1.