Detecting Speech Impairments from Temporal Visual Facial Features of Aphasia Patients

Moritz Einfalt, Rainer Lienhart
University of Augsburg
Universitaetsstr. 6a
86159 Augsburg, Germany
{*moritz.einfalt, lienhart*}*@informatik.uni-augsburg.de*

Matthew Lee, Lyndon Kennedy
FX Palo Alto Laboratory
3174 Porter Drive
Palo Alto, CA 94304, USA
{*mattlee, kennedy*}*@fxpal.com*

***Abstract*—We present an approach to detect speech impairments from video of people with aphasia, a neurological condition that affects the ability to comprehend and produce speech. To counter inherent privacy issues, we propose a cross-media approach using only visual facial features to detect speech properties without listening to the audio content of speech. Our method uses facial landmark detections to measure facial motion over time. We show how to detect speech and pause instances based on temporal mouth shape analysis and identify repeating mouth patterns using a dynamic warping mechanism. We relate our developed features for pause frequency, mouth pattern repetitions, and pattern variety to actual symptoms of people with aphasia in the AphasiaBank dataset. Our evaluation shows that our developed features are able to reliably differentiate dysfluent speech production of people with aphasia from those without aphasia with an accuracy of 0.86. A combination of these handcrafted features and further statistical measures on talking and repetition improves classification performance to an accuracy of 0.88.**

***Keywords*-facial features; speech diagnosis; medical assessment**

I. Introduction

Speech impairments are common symptoms for people with neurological conditions. Aphasia is one such neurological condition that affects a person's ability to understand or produce speech. Aphasia typically results from a stroke or other brain injury and can improve or worsen over time [1]. The degree and types of speech impairments in aphasia span a broad continuum, ranging from slightly dysfluent speech to severe limitations that only allow for a few words or utterances.

Assessing the abilities of people with aphasia is usually performed manually in direct interviews with doctors or therapists. The assessment itself can range from a broad classification of a patient's capabilities to a detailed analysis of symptoms based on interview transcripts [1]. Especially in the latter case, the time effort is enormous and may not even represent the person's abilities outside the clinical environment. Thus, there is an opportunity for automated assessment tools to track speech abilities more frequently over time and outside the clinical environment, enabling more effective tailoring of therapy for people with aphasia.

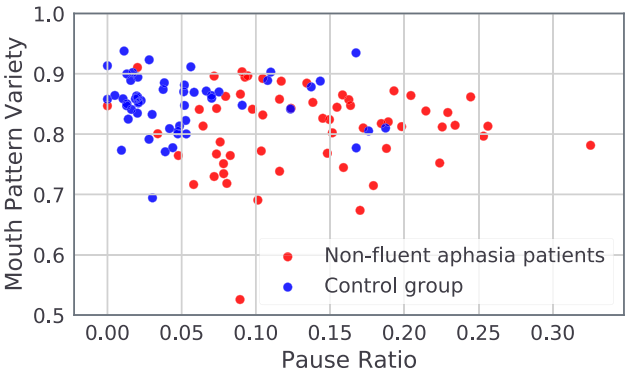

Figure 1. Example of our two best performing temporal features. In combination, they separate non-fluent aphasia patients from control group members in most cases.

Integrating assessment tools with everyday video calls (e.g. FaceTime) can provide more continuous and representative evaluations of speech impairments as they improve or worsen over time. A straightforward approach might focus on audio information to infer speech capabilities; however, it would require the system to listen in on what people actually say, which can raise concerns about privacy and confidentiality. In this paper, we propose a cross-media approach that only uses visual information to infer speech-related properties, while maintaining the person's need for privacy of their semantic speech content. This can increase the acceptance of automatic evaluation systems with possible applications in diagnosis and therapy [2], continuous assessment in video conferences or specific discourses with doctors, self-evaluation over time, or large scale medical studies.

Our approach is based on an initial registration of facial landmarks of people when they are talking. This is common for traditional facial vision tasks like face [3], emotion [4], [5] or gaze [6] recognition. Whereas the recognition targets for these tasks are typically related to visual facial properties, we instead specifically attempt to recognize speech properties. The task closest to our approach is visual speech

978-1-7281-1198-8/19/$31.00 ©2019 IEEE
DOI 10.1109/MIPR.2019.00026

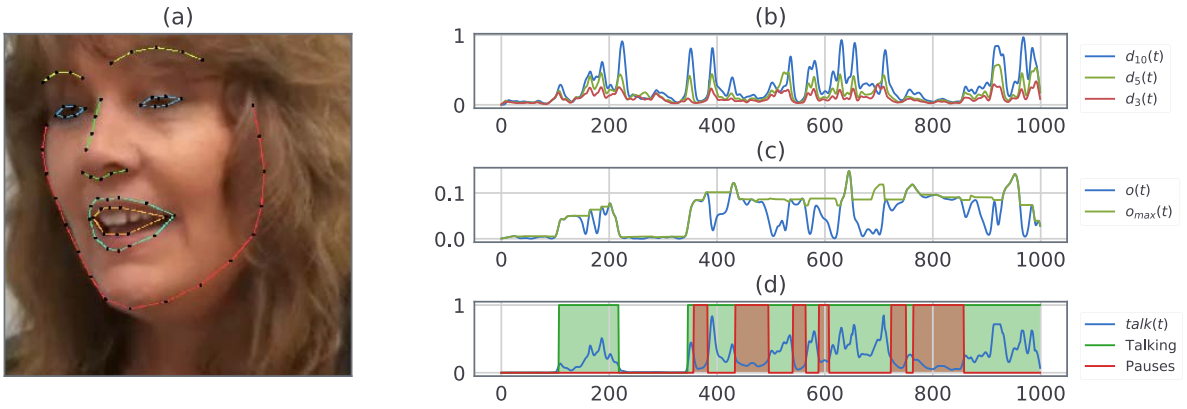

Figure 2. Overview of our basic features: (a) Video frame example from the *AphasiaBank* [1] dataset with detected facial landmarks. (b) Temporal mouth dissimilarity for different temporal windows. (c) Vertical mouth opening measure. (d) Combined talk score and the derived detection of talking intervals and pauses

recognition, with recent successes in automatic lip reading using DNNs [7]. Because we want to avoid extracting information on the semantic level, we focus on temporal features of facial motion to infer properties of speech patterns instead of the actual speech content. Our contributions are the development of temporal features for speech and pause detection, detection of repeating facial patterns, and the measure of overall facial pattern variety. We relate these features to actual speech-related symptoms of people with aphasia such as dysfluency, repetitive speech, and the use of a limited vocabulary. We report classification results using these features to distinguish different groups of people with aphasia from control group participants without aphasia, based on five minutes of interview recordings taken from the *AphasiaBank* [1] dataset.

II. Method

The basic idea of our approach is to analyze mouth shapes and mouth motion and to develop features related to actual speech properties. We use point detections outlining the mouth of a person in a video and compare the mouth shape over time. This reveals speaking turns as well as short pauses within speaking turns. Additionally, we group temporally sequential mouth shapes into mouth patterns and compare different patterns with each other to identify repeating patterns during talking. Finally, we measure the variety in mouth motion for a person by comparing it to a small vocabulary of observed patterns. Next, we describe each feature in detail.

A. Basic Features

The video material consists of recordings of participants (i.e. people with aphasia and a control group without aphasia) during interviews with fixed protocols. Even though the participant's face is either the only or the most prominent face in the recording, the camera viewpoint can vary from frontal to profile views of the participant's face. Our analysis begins with the registration of 2D landmarks on the person's face. We use a CNN-based approach following [8] to obtain a 70-point model for the characteristic facial points (see Figure 2a). Because many of the 70 facial points do not meaningfully change when the person is speaking, most of our analysis focuses on the mouth with its $M = 20$ points outlining the lips. We refer to this set of 2D mouth points at a specific point in time t in a video as:

$$\mathbf{m}_t = \begin{pmatrix} x_1 & x_2 & \cdots & x_M \\ y_1 & y_2 & \cdots & y_M \end{pmatrix}. \quad (1)$$

All video material is recorded at 30 frames per second, so we specify any point in time t by its frame index.

Our analysis of mouth configurations and their motion over time is based on a temporal similarity measure of a person's 2D mouth points and certain direct measurements of the mouth's opening. To compare how much the shape of a mouth changes over time, we measure the difference of two mouth configurations $\mathbf{m}_{t1}$ and $\mathbf{m}_{t2}$ based on their point-wise quadratic difference $\|\mathbf{m}_{t1} - \mathbf{m}_{t2}\|_2^2$. Any changes in the mouth configuration over time result from (a) body movement, (b) head movement or (c) inner-facial motion. To account only for the latter, we allow an arbitrary scaling s, 2D rotation $\mathbf{R}_\theta$ and translation $\mathbf{t}$ to map one of the compared mouth configurations as close as possible onto the other. We then use the remaining difference as the actual difference in shape. This is similar to the approach in [9] to capture body pose differences. In contrast to [4], we directly map the two mouth configurations we want to compare without an intermediate mapping onto a frontal template view of a mouth. Since we are later interested in the difference of temporally nearby mouth configurations in a video, we avoid the additional error induced by the intermediate face frontalization step [10]. Our approach is closely related to shape analysis with Procrustes methods [10], [11] and leads

to a mouth dissimilarity measure $msim$ with:

$$msim\left(\mathbf{m}_{t1}, \mathbf{m}_{t2}\right) = \min_{s,\theta,\mathbf{t}} \left\|\mathbf{m}_{t1} - s\mathbf{R}_\theta \mathbf{m}_{t2} + \mathbf{t}\right\|_2^2 . \quad (2)$$

$msim$ is inherently dependent on the scale of its first operand $\mathbf{m}_{t1}$. We therefore use the symmetric and scale invariant $msim_{\text{norm}}$ with:

$$msim_{\text{norm}}\left(\mathbf{m}_{t1}, \mathbf{m}_{t2}\right) = \frac{msim\left(\mathbf{m}_{t1}, \mathbf{m}_{t2}\right)}{s_{\mathbf{m}_{t1}}} + \frac{msim\left(\mathbf{m}_{t2}, \mathbf{m}_{t1}\right)}{s_{\mathbf{m}_{t2}}}, \quad (3)$$

where $s_{\mathbf{m}_t}$ is the average distance of any point in $\mathbf{m}_t$ from its center and acts as an estimate of its scale.

We can now use $msim_{\text{norm}}$ to measure inner-facial motion by comparing mouth configurations of the same person in a temporal window Δt. Because it is unclear which choice of Δt will result in the most informative measure, we use a collection of different temporal windows $\mathbf{w}$. This leads to our final temporal self-dissimilarity measure for the mouth $d_{\mathbf{w}}(t)$, defined as:

$$d_{\mathbf{w}}(t) = \sum_{\Delta t \in \mathbf{w}} msim_{\text{norm}}\left(\mathbf{m}_t, \mathbf{m}_{t+\Delta t}\right). \quad (4)$$

Our experiments revealed that short temporal windows $\mathbf{w} = \{2, 3 \ldots, 10\}$ are best suited to capture mouth changes during talking (see Figure 2b for examples).

B. Talking Detection

In order to infer different properties of someone's speech capabilities, the first step is to detect when someone is talking. (Certainly, talking can be easily inferred from the audio channel, but we restrict our approach to only visual to preserve privacy.) Since talking results in mouth movement, periods of talking reveal themselves as areas of high activity (or dissimilarity) in $d_{\mathbf{w}}(t)$. Note that $d_{\mathbf{w}}(t)$ not only captures mouth movement during talking but also (1) jitter in the point detections or missdetections and (2) changes due to out-of-plane rotations of the mouth, e.g. when nodding or shaking the head. To account for the former, we only consider time instances where all facial landmarks are detected with sufficient confidence. Registration errors from the latter effect are filtered out based on the observation that talking can only occur when the mouth is open at some point in the window. Thus, we measure the vertical distance between points on the upper and lower lips. Let $o(t)$ denote this vertical opening of the inner mouth, normalized by the scale of the complete face (analogous to the mouth scale $s_{\mathbf{m}_t}$). Since $o(t)$ is changing frequently during talking, we apply a closing operation to smooth over gaps of size up to $\Delta t = 50$ with a combined minimum and maximum filter and obtain $o_{\max}(t)$ with:

$$o_{\max}(t) = \min\left(\max_{t' \in [t-\Delta t, t]} o(t), \max_{t' \in [t, t+\Delta t]} o(t)\right). \quad (5)$$

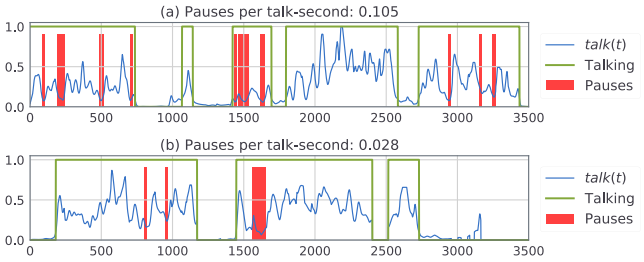

Figure 3. Detected pauses for (a) a person with aphasia and (b) a control group member.

Figure 2c depicts the result of this operation on a sample video. It fills gaps of a briefly closed mouth during talking while retaining sharp boundaries for longer periods of non-talking. Thus, a sufficiently high value of $o_{\max}(t)$ is a precondition for talking to be detected. Head motion without actual talking (or mouth opening) is filtered out, e.g. when a person is simply nodding. The final talk score is now given by $talk(t) = d_{\mathbf{w}}(t) \cdot o_{\max}(t)$, where $d_{\mathbf{w}}(t)$ and $o_{\max}(t)$ are maximum-normalized to $[0, 1]$. Finally, we apply a talk threshold τ_{talk} for a hard $\{0, 1\}$ assignment, remove very short talking intervals, and join closely adjacent talking intervals that are separated only by very short gaps. This represents our final talk instance detection (see Figure 2d).

C. Pause Frequency

One decisive symptom of impaired speech for people with aphasia is dysfluent speech, manifested as unintended pauses during talking. Developing a measure for pauses can lead to a direct measure of fluency. To detect pauses, we can again use the talk score $talk(t)$, apply a more restrictive threshold τ_{pause}, and register all areas of inactivity during the previously detected talk instances in Section II-B. Figure 2d shows an example for such pauses. Despite intended pauses being detected as well, we expect that the overall pause frequency is still related to a person's speech fluency. Figure 3 shows a qualitative example of the difference in pause frequency for a person with aphasia and a control group member.

D. Repetitive Patterns

Apart from dysfluency, one of the more noticeable speech symptoms of aphasia patients are frequent repetitions of utterances, words, or sentence fragments. These repetitions often occur when forming the subsequent word or when trying to correct the last word. We aim to find repetitions in the mouth motion that are related to speech repetitions. Even though repeating mouth motion may not be necessarily a direct indicator for repetition on the semantic level, statistics of visual repetition can still offer insight into repetition behaviors.

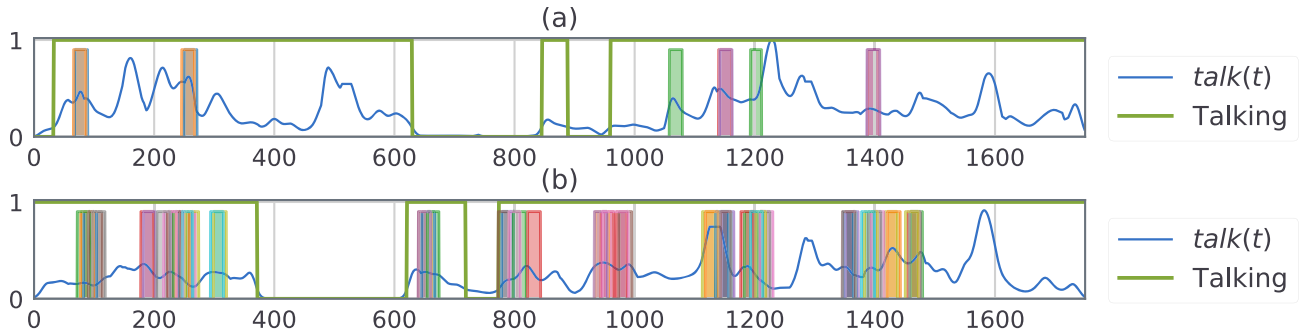

Figure 4. Repeating mouth patterns for (a) a person with aphasia and (b) a control group member. Different colors encode different mouth motion patterns. Pattern repetitions use the same color.

For our approach to detect visual repetitions of mouth motion, we define a pattern of mouth motion of length l around time t as $\mathbf{p}_t = (\mathbf{m}_{t-\lfloor l/2 \rfloor}, \ldots, \mathbf{m}_t, \ldots, \mathbf{m}_{t+\lfloor l/2 \rfloor})$. We compare two observed patterns of arbitrary length using Dynamic Time Warping [12], [13]: We transform $\mathbf{p}_{t1}$ into $\mathbf{p}_{t2}$ by either directly transforming a mouth configuration in $\mathbf{p}_{t1}$ into the respective configuration in $\mathbf{p}_{t2}$, or by allowing insertions or deletions in $\mathbf{p}_{t1}$. The cost for direct transformation operations is given by $msim_{\text{norm}}$ between the two transformed mouth configurations, maximum-normalized to $[0, 1]$. The more dissimilar both configurations are, the higher the cost. Insertion and deletion operations are always assigned the maximum cost of 1. Since the same mouth motion is not always performed at the same speed, temporal warping using insertions and deletions enables us to match similar patterns of different lengths. The overall pattern match cost is the cost sum of the optimal sequence of transformation operations.

To find possible repetitions, we first extract reference patterns around locally unique mouth configurations, i.e. maxima in $d_{\mathbf{w}}(t)$. We then search for matching patterns in the direct vicinity ($\pm$5s) that have a match cost below a certain threshold τ_{match}. Figure 4 shows examples for pattern matches for a person with aphasia that frequently repeats single words and a person without aphasia. Even though repeating patterns are found in both cases, the person with aphasia shows few but direct repetitions compared to the many highly interleaved repetitions for the person without aphasia. This hints that direct repetitions separated by no (or only a few) other patterns may be a good indicator for direct word repetitions. We therefore count the occurrences of direct repetitions and normalize by the total talking time to obtain a measure of (visual) repetitions per second.

E. Visual Vocabulary of Mouth Patterns

Instead of looking where in time certain patterns occur or repeat, we can also collect patterns over a complete video and assess their variety. This directly relates to the overall variety in mouth motion or expression and therefore - to a lesser extent - to the actual variety in speech. We would clearly expect a person capable of only expressing a few words or utterances to show less variety in mouth motion than an unimpaired person with a normal vocabulary.

To obtain such a measure of variety, we build a visual vocabulary of mouth patterns for each person by collecting a fixed number of patterns throughout a video and aggregating them using clustering with a fixed number of clusters. Only patterns that repeat at least once are selected. If not enough such patterns can be extracted, we select the missing ones randomly from the video. The representatives of all cluster centers form the vocabulary. For our approach we use k-medoids clustering with a predefined vocabulary size k. For a small k, we want to measure how well the vocabulary represents the complete variety of mouth motion of the person. We split the talking periods in the video into fixed-sized blocks of mouth motion. The blocks have the same length as the patterns in the vocabulary. Each block is now assigned the vocabulary element with the lowest pattern match cost from Section II-D. The idea is to reconstruct the complete mouth motion during talking by only concatenating the best fitting vocabulary elements. From this reconstruction, we measure the match (or reconstruction) cost between each block and its assigned vocabulary pattern and calculate the total reconstruction cost as the block-wise average. With limited variety in mouth motion, a small vocabulary suffices to describe the overall motion sufficiently well and leads to a good reconstruction. We compute the reconstruction cost for $k \in \{5, 10, \ldots, 50\}$ and use its mean as the final score of mouth motion variety.

III. Evaluation

We evaluate our developed features on the *Aphasia-Bank* [1] dataset of video recordings and transcripts of interviews with people with aphasia (APH) and control group participants (CTR). The aphasia patients are further classified based on their speech capabilities into fluent (FL) and non-fluent (NFL) speakers. While non-fluent patients show major impairments in physical speech production, fluent patients are usually able to talk more or less fluently but

the semantic content is often incorrect or incomprehensible. We removed a few videos of insufficient video quality and control group members below the age of 35 - an age group not represented in the dataset of people with aphasia. This resulted in a dataset of 163 CTR, 99 NFL and 111 FL videos. Because videos vary in length, we extract five minutes of direct discourse between the interviewer and the participants from each video for a fair comparison.

To evaluate the effectiveness of our developed features, we construct three binary classification tasks to discriminate the different groups of participants: CTR vs. NFL, FL vs. NFL and CTR vs. FL. Parameters for all speech features are optimized individually on a training set consisting of 45 CTR, 30 NFL, and 30 FL videos. The remaining videos are used as a test set for evaluation. Each classification task has a different ratio of positive to negative examples. In order to keep results across the different tasks comparable, we evaluate classification results using a *balanced accuracy* (ACC) measure. This is simply the mean accuracy on both classes, weighted by the fraction of examples in each class.

A. Individual Features

Table I shows the results when using each individual feature to directly classify participants on a video level. Each feature is treated as a likelihood score that a participant belongs to one class or the other. For the CTR vs. NFL task, i.e. discriminating normal and notably dysfluent speech, our inferred pause frequency measure performs best with an ACC of 0.86. This is not surprising, as pauses are inherently related to fluency of speech. The repetition frequency of mouth patterns reveals itself as not very informative with an ACC of only 0.64. To explain this result, Figure 5a shows the relationship between the detected pattern repetition frequency and the actual word repetition frequency derived from the video transcripts. It is obvious that there is no real correlation between both properties. Simply counting the number of directly repeating mouth patterns is therefore not suitable as an indicator for actual repetition in speech. As the third feature, the mouth pattern variety based on the concept of a visual vocabulary performs better, with an ACC of 0.72. Figure 5b shows that the visual vocabulary is able to identify parts of the aphasia patients with actual limited speech vocabulary by variety scores in the range [0.7, 0.8], which are hardly observed for the control group. However, it is not well-suited as a single discriminating feature to distinguish the two participant groups on its own, but in combination with other features (see Figure 1), it still contains viable information.

The relative ordering of feature performance stays the same for the FL vs. NFL task, but the overall performance drops significantly with only 0.69 ACC for the pause frequency. This is somewhat to be expected as the ground truth separation into fluent and non-fluent aphasia patients is based on a subjective impression of the interviewer. Both

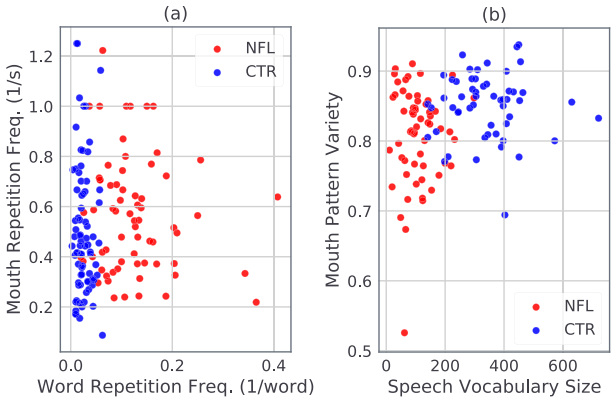

Figure 5. Relationship of developed features and speech properties: (a) Detected mouth pattern repetitions and actual word repetitions from the transcript. (b) Mouth pattern variety and actual vocabulary size. Only the latter shows informativeness.

groups still share symptoms of varying degrees and are therefore much more difficult to distinguish.

For the final task CTR vs. FL, performance slightly recovers again. Most notably, the pause frequency and the mouth pattern variety perform very similar with 0.74 and 0.70 ACC each. Since highly non-fluent participants are not present in this task, the pause frequency as a direct measure is not necessarily the most informative feature.

B. Feature Combination

Different aphasia patients may vary in the types and degrees of speech symptoms, so a single feature extracted from the interview videos might not be optimal in differentiating different types of aphasia in general. We therefore additionally examine if the combination of features within an arbitrary classifier outperforms the classification using each feature separately. To this end we use a random forest classifier on our developed features and apply it to the same three classification tasks. We choose its hyper-parameters based on the training set and report the results from a 5-fold cross-validation on the test set in Table I. With the exception of the CTR vs FL task, the performance of the classifier seems to be capped by the best performing single feature. Since randomized classifiers in general benefit from a large set of features to draw from, the performance using only the three specifically designed and parameterized features is not necessarily optimal. However, many other features can be extracted from our concepts of talk and pause detection, pattern repetition and pattern variety. We repeat the experiment by including further statistical measures which are not directly related to specific symptoms. Our choice includes the average length of talk intervals and pauses as well as the average duration between pattern repetitions. Extending the feature set leads to gains in all three classification tasks compared to the single feature performance, with an ACC of up to 0.88. Therefore, adding more statistical features built

	Pause Freq.	Repetition Freq.	Mouth Pattern Variety	Feature Combination	Extended Features
CTR vs. NFL	**0.86**	0.64	0.72	0.86	**0.88**
FL vs. NFL	**0.69**	0.59	0.60	0.68	**0.69**
CTR vs. FL	**0.74**	0.64	0.70	**0.76**	**0.76**

Table I
BALANCED ACCURACY ON THE THREE CLASSIFICATION TASKS USING INDIVIDUAL FEATURES, FEATURE COMBINATIONS WITH A RANDOM FOREST CLASSIFIER, AND AN EXTENDED FEATURE SET WITH ADDITIONAL TALK AND REPETITION STATISTICS.

on our concepts of talk detection and repeating patterns leads to an increase in discriminative power, but at the cost of losing the semantic interpretation of the individually developed features. Thus, computational approaches such as ours may be able to identify subtle patterns in behaviors related to aphasia that traditional therapists may not normally be able to detect.

IV. CONCLUSION

We presented a cross-media approach to infer speech impairments of people with aphasia based on visual facial features alone, without the need to listen to what they say. Based on detections of facial landmarks, we applied techniques from shape analysis and sequence warping to develop methods for detecting periods of talking and pauses as well as repetitions in facial temporal patterns. We additionally measured the variety of facial patterns based on a visual pattern vocabulary. Our evaluation showed that measures of pause frequency and the variety of mouth patterns are useful in differentiating people with and without non-fluent aphasia. Combining individual features together, along with additional statistics on repetition and talking, resulted in an overall balanced accuracy of 0.88 for distinguishing people with and without non-fluent aphasia. In the future, we aim to include body movement such as head motion and hand gestures into our analysis to detect gesturing as a replacement for missing speech capabilities, as well as, expanding to other neurological conditions such as dementia.

ACKNOWLEDGMENT

This research was conducted during Moritz Einfalt's internship at FX Palo Laboratory. He thanks the colleagues from FXPAL for the collaboration and for providing an open and inspiring research environment.

REFERENCES

[1] B. MacWhinney *et al.*, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[2] L. R. Cherney, "Oral reading for language in aphasia (orla): Evaluating the efficacy of computer-delivered therapy in chronic nonfluent aphasia," *Topics in Stroke Rehabilitation*, vol. 17, no. 6, pp. 423–431, 2010.

[3] K. Niinuma *et al.*, "Automatic multi-view face recognition via 3d model based pose regularization," in *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sept 2013, pp. 1–8.

[4] J. Wang *et al.*, "Video-based emotion recognition using face frontalization and deep spatiotemporal feature," in *First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, May 2018, pp. 1–6.

[5] A. R. Babu *et al.*, "Facial expressions as a modality for fatigue detection in robot based rehabilitation," in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference (PETRA)*. ACM, 2018, pp. 112–113.

[6] A. C. Varchmin *et al.*, "Image based recognition of gaze direction using adaptive methods," in *Gesture and Sign Language in Human-Computer Interaction*. Springer Berlin Heidelberg, 1998, pp. 245–257.

[7] S. Petridis *et al.*, "End-to-end visual speech recognition with lstms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2592–2596.

[8] Z. Cao *et al.*, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] R. Lienhart, M. Einfalt, and D. Zecha, "Mining automatically estimated poses from video recordings of top athletes," *International Journal of Computer Science in Sport (IJCSS)*, vol. 17, no. 2, pp. 94 – 112, 2018.

[10] B. Martinez *et al.*, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, 2017.

[11] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 285–339, 1991.

[12] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," in *Readings in speech recognition*. Elsevier, 1990, pp. 159–165.

[13] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 285–289.