Unsupervised estimation of speech rhythm, articulation, and phonation quality for clinical applications

Yishan Jiao[1], *Visar Berisha*[1,2], *Julie Liss*[1]

[1]Department of Speech and Hearing Science
[2]School of Electrical, Computer, and Energy Engineering
Arizona State University
{yjiao16, visar, julie.liss}@asu.edu

Abstract

Automatic speech signal analysis is useful in many fields, such as speech recognition, speaker identification, clinical diagnosis and so on. This paper describes a series of methods for estimating the quality of three elemental aspects of speech - rhythm, articulation and phonation. The principal aim of this work is to develop robust methods for detecting unusual changes in speech for clinical applications. Most of the existing work in this area relies on supervised learning with labeled pathological speech. In contrast, we focus on either completely unsupervised algorithms or algorithms trained only on healthy speech datasets. For rhythm, we describe a robust speaking rate (SR) estimation algorithm which provides online SR trajectories over time. For articulation, we describe an unsupervised measure of articulation precision and a representation of likely phonological errors made by a speaker. Estimation of phonation quality is our next goal. This paper concisely describes some of the methods and key results associated with this work.

Index Terms: pathological speech, rhythm, articulation, phonation quality, unsupervised

1. Introduction

According to World Health Organization (WHO), the proportion of people aged 60 or more is estimated to almost double from about 12% in 2015 to 22% in 2050. Among this population, over 20% suffer from a neurological disorder that impacts their ability to produce clear and intelligible speech, leading to communicative disabilities. Recent results suggest that subtle changes in speech and language precede more obvious symptoms [1]. Furthermore, for those already diagnosed and seeking clinical intervention, the number of certified speech-language pathologists (SLPs) who help treat these problems is limited with waiting time of seeing a specialist exceeding 4 months in the US. Therefore, there is an urgent need for technology that can automatically detect subtle changes in speech for early detection of neurological disorders or, for those already diagnosed, can augment traditional clinical care in order to lighten the load for SLPs.

This is a recognized problem that has attracted growing attention in the field of speech signal processing. Most of the previous studies have focused on classification or detection of speech disorders from speech signals [2]. However, the diagnosis of the disease is only the first step of therapy and these algorithms provide limited information regarding rehabilitation. Other studies have tried to estimate the intelligibility (percent words correct) of pathological speech [3]. Although improving intelligibility is critical in clinical care, it is an abstract concept that is impacted by a number of factors related to speech production: for example, a decrease of intelligibility can result from imprecise articulation, abnormal prosody, pathological phonation, or combinations of these. Clinicians must perceptually decompose speech along these dimensions and detect anomalies along each in order to develop personalized treatment plans. Apart from these, there are a few studies that propose computer-aided systems for clinical use [4]. However, most of the system were assembled by using existing algorithms developed without considering pathological applications, or need some prior knowledge to adapt.

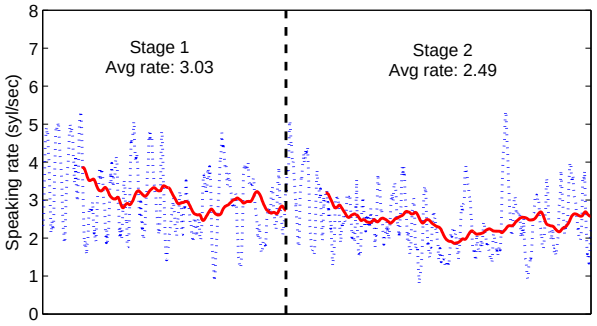

Figure 1: Results of speaking rate estimated on longitudinal dysarthric speech. Blue line is the estimated rate on each 1 second speech with a 0.1 second shift. The red line is the average speaking rate of each 10 seconds speech.

In contrast to previous work, we aim to develop a suite of tools for evaluating pathological speech along dimensions that directly impact intelligibility. This helps provide actionable information to clinicians regarding intervention. Our approach is to develop a series of fast and reliable algorithms from different aspects of speech by only using patients' speech signals. The applications are various. The algorithms can be used as a part of a partially automated computer-aided system for intervention; to detect subtle, but remarkable, changes in patients at risk for a neurological disorder; or as a reference for clinicians and patients to monitor and track disease progress.

2. Method and Result

2.1. Online speaking rate estimation

Most of the SR estimation algorithms in the literature estimate the rate by counting the number of syllables in a speech segment using energy-based features followed by thresholding to detect

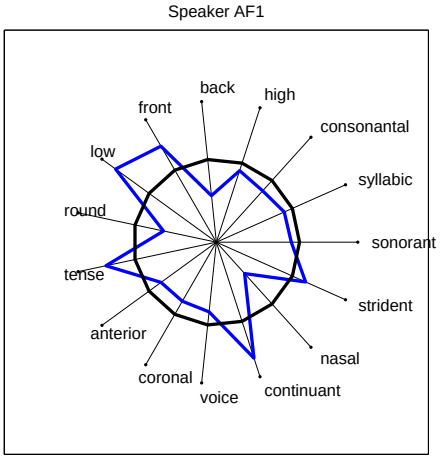

Figure 2: A spiderplot shows the difference between a dysarthric speaker and the average of healthy groups in 15 phonological groups.

syllables. However, the models are usually less robust due to the heuristic parameters used in the algorithms. To avoid these, we proposed to use the envelope modulation spectrum (EMS) combined with the statistics of acoustic features from a speech segment, and to train a recurrent neural work to predict the SR from these features in 1-sec intervals [5][6]. The resulting algorithms outperform other state-of-the-art methods on different types of healthy and dysarthric speech. For example, Figure 1 shows the SR estimation results of a dysarthric speaker reading the same passage across 2 stages of the disease. The algorithm makes it possible to view the change in SR over time in completely unsupervised fashion.

2.2. Articulation entropy[1]

Articulation precision is one of the critical factors influencing speech intelligibility. A common measure of articulation precision is the vowel space area (VSA), which is the area of the quadrilateral spanned by the first and second formants of the 4 corner vowels. While this is a commonly used metric in pathological speech analysis, it only considers voiced segments of speech while discarding the unvoiced period which constitute 1/3 of the speech. Therefore, we proposed a more general measure we call 'articulation entropy'. We extract 13th-order mel-spectrum features with cubic root compression (MelRoot3) [7] from every frame of speech sample and calculate the Rényi entropy of the distribution composed of these features from a given speech sample. The entropy is calculated using a non-parametric entropy estimation algorithm [8]. We posit that larger values in entropy will correspond to better articulation precision. The advantages of articulation entropy compared with VSA are: (1) It is completely unsupervised and requires no manual labeling of speech. (2) It can be estimated on speech samples of varying length. To evaluate the performance of this measure, we have conducted several experiments on different datasets, such as English dysarthric speech, the speech of patients with cochlear implants, and Mandarin dysathric speech. On all data sets, the proposed algorithm successfully tracks perceived perception of articulation precision.

[1]The papers describing articulation entropy and the following phonological analysis interface are currently in progress.

2.3. Articulation of different phonological categories

There is no shortage of acoustic features used by speech signal processing experts to describe speech production. However, these features are rarely used by clinicians because of their lack of interpretability. Instead, phonological features based on class, manner and place of articulation, are more comprehensible to clinicians. We have trained a recurrent neural network to map acoustic features into 15 phonological categories using healthy speech [9]. The trained model was used to evaluate dysarthric speech by algorithmically analyzing differences between healthy and dysarthric phonological feature statistics and the result was presented using a spider plot. In Figure 2 we show a sample plot that describes the articulation quality of a dysarthric speaker. For each phonological group, the closer a speaker is to the circle, the more precise their ability to produce sounds from that category. This representation provides clinicians with an interpretable means of precisely characterizing articulation.

3. Conclusion and Future Work

In this PhD project, our goal is to develop a suite of automatic speech analysis tools for clinical use. We have proposed robust algorithms for estimating rate and articulation precision and will develop the same for estimating phonation quality of speech. The proposed algorithms have been successfully validated on different data sets.

4. References

[1] A. P. Vogel, C. Shirbin, A. J. Churchyard, and J. C. Stout, "Speech acoustic markers of early stage and prodromal huntington's disease: A marker of disease onset?" *Neuropsychologia*, vol. 50, no. 14, pp. 3273–3278, 2012.

[2] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, vol. 1. IEEE, 2002, pp. 182–183.

[3] K. C. Hustad, "Estimating the intelligibility of speakers with dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 3, pp. 217–228, 2006.

[4] A. F. Seddik, M. El Adawy, and A. I. Shahin, "A computer-aided speech disorders correction system for arabic language," in *2013 2nd International Conference on Advances in Biomedical Engineering*. IEEE, 2013, pp. 18–21.

[5] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Online speaking rate estimation using recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5245–5249.

[6] Y. Jiao, V. Berisha, M. Tu, and J. Liss, "Convex weighting criteria for speaking rate estimation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 9, pp. 1421–1430, 2015.

[7] M. Tu, X. Xie, and Y. Jiao, "Towards improving statistical model based voice activity detection," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] A. O. Hero, B. Ma, O. J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE signal processing magazine*, vol. 19, no. 5, pp. 85–95, 2002.

[9] A. Asaei, M. Cernak, and H. Bourlard, "On compressibility of neural network phonological features for low bit rate speech coding," in *Proc. of Interspeech*, 2015, pp. 418–422.