Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Generative Model-Driven Feature Learning for dysarthric speech recognition

N. Rajeswari [a,*], S. Chandrakala [b]

[a] Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India
[b] Department of Computer Science and Engineering, Rajalakshmi Engineering College, Rajalakshmi Nagar, Thandalam, Chennai, India

ARTICLE INFO

Article history:
Received 29 December 2015
Received in revised form
11 April 2016
Accepted 29 May 2016
Available online 6 June 2016

Keywords:
Generative Model-Driven Feature Learning
Dysarthric speech recognition
Support vector machine
Varying length sequences
Feature vector representation

ABSTRACT

Recognition of speech uttered by severe dysarthric speakers needs a robust learning technique. One of the commonly used generative model-based classifiers for speech recognition is a hidden Markov model. Generative model-based classifiers do not do well for overlapping classes and due to insufficient training data. Dysarthric speech is normally partial or incomplete that leads to improper learning of temporal dynamics. To overcome these issues, we focus on learning features for dysarthric speech recognition that involves recognizing the sequential patterns of varying length utterances. We propose a Generative Model-Driven Feature Learning based discriminative framework that maps the sequence of feature vectors to fixed dimension vector spaces induced by the generative models. The discriminative classifier is built in that vector space. The proposed HMM-based fixed dimensional vector representation provides better discrimination for dysarthric speech than the conventional HMM. We examine the performance of the proposed method to recognize the isolated utterances from the UA-Speech database. The recognition accuracy of the proposed model is better than the conventional hidden Markov model-based approach.

1. Introduction

Dysarthria is a kind of motor speech disorder caused by neurological injury to the central or the peripheral nervous system. Speech subsystems such as respiration, phonation, resonance, prosody, and articulation can be affected, leading to the detriments in intelligibility (how well the speaker's speech is understood), audibility, naturalness, and potency of vocal communication. This kind of disorder is caused by a stroke, muscular dystrophy, brain injury, tumours, Parkinson's disease, or Huntington's disease, or multiple sclerosis.

Some dysarthric speech characteristics [1–3] are mono pitch, harsh voice, vowel distortions, strained-strangled vocal quality. Due to these characteristics, the pronunciation often suffers from the following limitations: the rate of the dysarthric speech is lower; there is no consistency in pronunciation; pronunciation varies due to fatigue; speaking

* *Corresponding author at*: Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, Tamil Nadu 602117, India.
E-mail addresses: raji@svce.ac.in, rajeswari.natarajan@gmail.com (N. Rajeswari).

http://dx.doi.org/10.1016/j.bbe.2016.05.003
0208-5216/

rate which is so important for speech recognition is slow. Automatic Speech Recognition (ASR) systems can be very helpful to recognize speech of dysarthric individuals because the people affected from dysarthria suffer from a severe neurological disorder. Conventional ASR systems for non-dysarthric speakers have not yielded good recognition rates for speakers suffering from dysarthria [4,5]. Thus, developing a robust ASR model that is solely meant for speakers with dysarthria is important.

One of the commonly used generative model-based classifiers for classifying the sequential patterns of varying length are hidden Markov models (HMMs). The sequence of feature vectors extracted from speech utterances is sub-word unit or word unit. These units of speech are characterized by a short duration of about 100 to 500 ms. A class-specific HMM is built using the sequence of feature vectors extracted from all the training utterances of a class. The class label for a test utterance is assigned for which the log-likelihood score is maximized. Baum–Welch method, an expectation maximization based method to find the maximum likelihood (ML), or maximum a posteriori (MAP) method is used for estimating the parameters of an HMM of each class. HMM-based classifiers using the ML and MAP methods for parameter estimations are robust only if enough training data is available or if the utterances are normal speech and not the partial or impaired speech. Generative model-based classifiers are not suitable for classifying the data of overlapping classes because a model is built for each class using the data belonging to that class alone. Overlapping of dysarthric speech arises due to similar characteristics between the classes and due to partial speech with missing contents. It is also important to design a discriminative model-based approach for classifying overlapping utterances of impaired speech as the conventional HMM-based method is a non-discriminative training based method.

We aim at learning a fixed dimensional vector representation for dysarthric speech recognition that involves recognizing the sequential patterns of varying length utterances. Learning features are significant for complex tasks such as dysarthric speech recognition as dysarthric speech is normally incomplete due to imprecise articulation, hyper-nasality and bilateral weakness. We propose a Generative Model-Driven Feature Learning method that uses the generative model driven data representation with a discriminative model-based classifier. We use the log-likelihood scores and transition probability scores obtained from HMMs as features.

The rest of this paper is organized as follows: Section 2 presents a review of different approaches for dysarthric speech recognition task. In Section 3, we describe the conventional hidden Markov model-based approach for modelling sequences of varying length feature vectors. Generative Model-Driven Feature Learning method for DSR is presented in Section 4. Studies on comparison of conventional HMMs with that of the proposed method are presented in Section 5. The conclusion is presented in Section 6.

2. Background

Dysarthric speech recognition systems proposed earlier are based on hidden Markov models, support vector machines (SVMs), Artificial Neural Networks (ANNs) or hybrid framework that uses both generative and discriminative learning-based algorithms. The ASR system [6] used the speech of three dysarthric speakers (one female, two males) and one control speaker. All three speakers possessed the symptom of low intelligibility. Hidden Markov model and support vector machine were used to develop an isolated digit recognition system for three subjects. HMM-based recognizer was successful for two subjects but failed for the subject whose utterances had missing consonants. SVM-based recognition system, assuming fixed word length proved to be good for two subjects, but failed for the subject with the stutter. An assisting speech-enabled system [5] used hidden Markov model to recognize dysarthric speech using Mel Frequency Cepstral Coefficients (MFCC) as features and Perceptual Evaluation of the Speech Quality (PESQ) [7] measure to assess the quality of speech. Experiments used the speech of four male American speakers with severe dysarthria from Nemours database [8], and two French speakers from Acadian corpora (collected by them) with sound substitution disorders (SSD). It has been reported that the best recognition rate can be got when the Hamming window size is greater than 25 ms.

Another HMM-based approach [9] was used to model the large variability of speech for isolated word recognition. Out of two results, the first result, showed that ergodic HMM outperformed a standard left-to-right (Bakis) model structure. The second result found that, Automated transition clipping acoustics enhanced recognition. The vocabulary included 10 digits and 196 common words. A speaker-adaptive phoneme-based system, and a speaker-dependent, whole-word pattern-matching system [10] projected the feasibility of using speech recognizer as a text input method for dysarthric speakers with different intelligibility levels. Two male and two female dysarthric participants were used for evaluation. One had a mild severity level, one had a moderate severity level and two dysarthric participants suffered from severe dysarthria. All the participants were Swedish dysarthric speakers. Speakers with severe dysarthria recognized repeated texts, despite low speech intelligibility scores.

The STARDUST (Speech Training And Recognition for Dysarthric Users of ASsistive Technology) project developed an HMM-based robust speech recognizer for three dysarthric speakers whose intelligibility is unknown. The recognizer used MFCC features for a ten-word vocabulary. The intelligibility of the dysarthric speakers is unknown. Further experiments [11] used the speech of eight low intelligibility dysarthric subjects (2 female and 5 male). Speech training using computer generated auditory and visual feedback reduced the inconsistent production of key phonetic tokens. An HMM-based system [12] was studied, on how best the dysarthric speech can be recognized by the continuous speech recognizer trained on non-dysarthric speech. Speaker independent and speaker dependent HMMs trained using non-dysarthric speech of two Dutch male speakers and were evaluated using the speech of two Dutch male dysarthric speakers. ASR of dysarthric speech was within the reach for higher perplexity tasks, even when the speaker speaks at a slower rate.

An Artificial Neural Network (ANN)-based model [13] used two multi-layer neural networks to recognize dysarthric speech of a male speaker with cerebral palsy for a ten word