Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 228–242

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Hierarchical ANN system for stuttering identification☆

Izabela Świetlicka[a,*], Wiesława Kuniszyk-Jóźkowiak[b], Elżbieta Smołka[b]

[a] *Department of Physics, University of Life Sciences, Akademicka 13, 20-950 Lublin, Poland*

[b] *Laboratory of Biocybernetics, Institute of Computer Science, Maria Curie-Sklodowska University, 1 Maria Curie-Sklodowska sq, 20-031 Lublin, Poland*

Received 16 June 2011; received in revised form 20 March 2012; accepted 15 May 2012
Available online 24 May 2012

Abstract

The presented work covers the issue of applying neural networks to the recognition and categorization of non-fluent and fluent utterance records. Speech samples containing three types of stuttering episodes (blocks before words starting with plosives, syllable repetitions, and sound-initial prolongations) were applied. The proposed system, built with hierarchical neural network framework, was used and then evaluated with respect to its ability to recognize and classify disfluency types in stuttered speech. The purpose of the first network was to reduce the dimension of vector describing the input signals. The result of the analysis was the output matrix consisting of neurons winning in a particular time frame. This matrix was taken as an input for the next network. Various types of MLP networks were examined with respect to their ability to classify utterances correctly into two, non-fluent and fluent, groups. Good examination results were accomplished and classification correctness exceeded 84–100% depending on the disfluency type.

Keywords: Artificial neural network; Kohonen network; Multilayer Perceptron; Classification; Stuttering

1. Introduction

Speech is considered to be a very effective and developed means of communication, which delivers the set of information concerning not only the main content of the statement but also the speaker's emotional condition, his or her age, intentions and many other features, that, seemingly, do not have any connection with the statement. On their basis, however, the relations between the listener and the speaker are built as well as both opinions and judgements about the speaker are formed. The main aim of the speech process is to send and receive messages in the linguistic form. Sometimes, however, the message might be affected by physiological or psychological factors, becoming incoherent for the listener. One possible reason disrupting the process of conveying information is stuttering.

Despite a large body of research, our understanding of stuttering still lacks consensus. The description of stuttering, as well as the ability of human judges to agree consistently on what constitutes an episode of stuttering, leads to many contradictions (Brundage et al., 2006; Cordes, 2000). This translates into inadequate measurement of stuttering severity and has an impact on therapy outcome measures. Therefore, creating an artificial system for measurement of disfluency

☆ This paper has been recommended for acceptance by "Simon King, Ph.D.".

* Corresponding author. Tel.: +48 81 445 69 05; fax: +48 81 533 35 49.
E-mail address: izabela.swietlicka@up.lublin.pl (I. Świetlicka).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.05.003

in speech might offer a solution to such problems (Howell et al., 1997b, 1999; Czyżewski et al., 2003; Gelzinis et al., 2008).

Nowadays, applications that allow for the identification of disturbances occurring in speech in a more or less automatic way are formed mainly on the grounds of acoustic and frequency features. Artificial recognition and disfluency identification are considered to be complicated and complex but benefits of creating such a system are obvious e.g. in the area of respiratory system diseases (Behroozmand and Almasganj, 2007; Gelzinis et al., 2008; Wszołek et al., 2001) or stuttering (Czyżewski et al., 2003; Fragopanagos and Taylor, 2005; Gas et al., 2004; Shao and Barker, 2008). Many attempts at research in the domain of speech classification and processing have been undertaken however suggested solutions concern relatively 'clear' speech (Wouhaybi and Al-Alaoui, 1999; Ververidis and Kotropoulos, 2006; Trentin and Giuliani, 2001; Siniscalchi et al., 2006; Shi et al., 2006; Nicholson et al., 2000; Kotti et al., 2008; Kocsor et al., 2000; Hosom, 2003; Fragopanagos and Taylor, 2005) while only few attempts have been made to examine disordered speech. The majority of people (not only people who stutter) have some problems with fluency while speaking, so automatic speech recognition (ASR) applications need to be equipped with solutions to improve their performance. Existing approaches to dealing with disfluencies in typical (not stuttered) speech or noisy environment cover mainly MFCC, LPC, SVMs, HMMs and ANNs (Tadeusiewicz and Ogiela, 2004; Ogiela et al., 2005, 2006, 2008; Ezeiza et al., 2011; Glasberg and Moore, 1990; Patterson et al., 1995) application.

Disfluency, especially stuttering, is usually accompanied by changes in breathing, phonation, and articulation, affects accent, speech rate and rhythm (Bloodstein, 1995; Tarkowski, 1999) which presents many problems for speech recognition applications. To build systems supporting disfluent speech recognition, various mathematical methods such as HMMs (Walles and Hansen, 1996; Wiśniewski et al., 2006), fuzzy logic (Suszyński et al., 2003), correlation function (Suszyński et al., 2005), genetic algorithms (Behroozmand and Almasganj, 2007) and many others have been used, but it is the projection of human aural perception system that seems to be the most promising approach to the subject. The artificial neural networks (ANN) that are known as a simplified model of biological nervous system (Katagiri, 2000; Bishop, 1995) allow for such observation of the non-fluent speech analysis. That is why the growing trend in the ANN or ANN and HMMs hybrid applications could be observed in such areas as speech and speaker recognition (Farrell, 2000; Farrell et al., 1994; Hadjitodorov et al., 1997; Leung et al., 2007; Trentin and Giuliani, 2001) and classification (Hosom, 2003; Cosi et al., 2000; Kocsor et al., 2000; Lee et al., 1998) or feature extraction (Katagiri, 2000; Gemello et al., 2007; Fritsch et al., 2000; Yegnanarayana and Narendranath, 2000; Shao and Barker, 2008; Uncini, 2003). The Multilayer Perceptron (MLP) and Radial Basis Function networks (RBF) as well as recurrent and fuzzy networks frequently occur in the automatic speech recognition (ASR) process (Chen et al., 1996; Farrell, 2000; Leung et al., 2007; Schuster, 2000). In addition, Kohonen's Self-organizing Maps are considered to be very useful as far as data compression and transformation are concerned (Kotti et al., 2008). Thanks to their features – input data generalization, distinguishing between very similar signals, etc., ANNs are widely used in the area of non-fluent speech recognition, especially in classification (Czyżewski et al., 2003; Howell and Sackin, 1995; Howell et al., 1997b, 1997a; Nayak et al., 2005), speech quality assessment (Geetha et al., 2000), quite often with the use of topological maps (Leinonen et al., 1993; Wszołek and Tadeusiewicz, 2000). The diversity of ANN applications at many stages of the speech recognition process indicates that artificial neural networks are effective in both signal recognition and classification.

Artificial neural networks are widely applied in the field of classification. One of the main reasons of the ANN popularity in the area is the fact that, in contrast to traditional statistical methods, networks adjust data without the necessity of defining any additional function or distribution of input variables. They are also able to determine the probability of an element belonging to the group that permits the use of the ANN application as a posteriori probability estimators for some specified objects (Hung et al., 1996; Kline and Berardi, 2005; Bernardi and Zhang, 1999). For binary classification ($j=2$), a posteriori probability that the object $\mathbf{x}$ belongs to the group w_1 takes the following form (1):

$$P(w_1|\mathbf{x}) = \frac{p(\mathbf{x}|w_1)P(w_1)}{p(\mathbf{x}|w_1)P(w_1) + p(\mathbf{x}|w_2)P(w_2)} \tag{1}$$

where $P(w_1)$ or $P(w_2)$ is the a priori probability that the random object $\mathbf{x}$ belongs to the group w_1 or w_2, $P(w_1|\mathbf{x})$ the a posteriori probability that the random object $\mathbf{x}$ belongs to the group w_1, $P(\mathbf{x}|w_1)$ is the sample distribution for the class w_1.

Formula (1) shows that there is a possibility of such selection of neuron parameters that the probability of the occurrence of classification error is minimized (Hassoun, 1995; Ripley, 1996).