Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 122–135

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Automatic detection of stridence in speech using the auditory model☆

Ružica Bilibajkić[a,*], Zoran Šarić[a], Slobodan T. Jovičić[a,b], Silvana Punišić[a], Miško Subotić[a]

[a] *Life Activities Advancement Center, Gospodar Jovanova 35, 11000 Belgrade, Serbia*
[b] *School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11000 Belgrade, Serbia*

Received 17 September 2014; received in revised form 22 April 2015; accepted 21 August 2015
Available online 21 September 2015

Abstract

Stridence as a form of speech disorder in Serbian language is manifested by the appearance of an intense and sharp whistling. Its acoustic characteristics significantly affect the quality of verbal communication. Although various forms of stridence manifestation are successfully diagnosed by speech therapists, there is a need for the automatic detection and evaluation of stridence. In this paper, an algorithm for stridence detection using Patterson's auditory model is presented. The algorithm consists of three processing stages. In the first stage spectral analysis and masking effects are applied using Paterson's auditory model. In the second stage a contour of spectral peaks that best fits characteristic features of the stridence is selected in the time-frequency (TF) representation of the signal obtained by Patterson's auditory model. In the third stage hypothesis testing is performed with three decisions: D_0 – no stridence, D_1 – stridence, and D_2 – unable to decide. The reliability of stridence detection is tested on the speech corpus of 16 speakers without stridence (with correct speech), 16 speakers without stridence but with some other speech sound disorders, and 16 speakers with stridence. Test results show high correspondence of subjective measures and automatic detection.

Keywords: Speech pathology; Stridence; Pathology detection; Auditory model

1. Introduction

Fricatives and affricates are groups of phonemes with common problems in pronunciation. Fricatives are articulated by forcing air through a narrow constriction in the vocal tract, resulting in a steady friction noise, while affricates are articulated by forming a complete constriction, which release is accompanied by the friction noise. Like affricates, oral stops are also formed by a complete constriction situated in the vocal tract, but they differ in the manner of the release of the confined air. Oral stops are characterized by abrupt release, while affricates are released with the constricted

☆ This paper has been recommended for acceptance by Katrin Kirchhoff.

* Corresponding author at: Life Activities Advancement Center, Gospodar Jovanova 35, 11000 Belgrade, Serbia. Tel.: +381 693086788/113208513; fax: +381 112624168.

E-mail addresses: r.bilibajkic@add-for-life.com (R. Bilibajkić), sariczoran@yahoo.com (Z. Šarić), jovicic@etf.rs (S.T. Jovičić), silvanapunisic@hotmail.com (S. Punišić), ifp2@ikomline.net (M. Subotić).

http://dx.doi.org/10.1016/j.csl.2015.08.006
0885-2308/

vocal tract creating the friction of the air passing through. Our study is focused on irregularities in the articulation of fricatives and affricates related to the friction control and synchronized movements of articulators. In the case of normal speech and language development, fricatives and affricates are acquired at the age of 6–7. During the developmental phase, irregularities in the articulation of some fricative speech sounds may occur. These irregularities may be the result of the normal speech development, or may be caused by pathology in sound production. Types of deviations occur depending on the pattern of speech sound groups, and can be language dependent. These deviations manifest themselves as different types of sigmatism, duration, intensity and friction quality impairment (Jovičić et al., 2010). Sigmatism or lisp is an articulatory disorder characterized by defective sibilant sound. Speech therapists recognize five types or lisps: frontal or interdental, lateral, occluded, nasal and strident (Riper and Erikson, 1996). In the case of the Serbian language, strident lisp or stridence is one of the specific forms of deviations that occur in the articulation of fricatives and affricates.

Stridence is an acoustic phenomenon that is generated in the mouth when the position of the tongue in relation to the palate and teeth is irregular. This position creates constrictions of various forms. When the airflow reaches a certain speed while passing through constrictions a tone of a certain frequency or a very strong narrowband resonant noise is generated. These sounds are generated simultaneously with the pronunciation of a speech sound, typically fricatives and affricates, and they change the acoustic characteristics of the target phoneme. Perceptually, the stridence is experienced as an unpleasant, whistling, squeaky or coarse sound that influences the quality of the pronounced phoneme.

Stridence is also considered to be an abstract phonological distinctive feature (Chomsky and Halle, 1968) and is a phonological feature of many languages. In addition, the term strident lisp is used in order to describe the sibilants characterized by piercing, whistling sounds (Riper and Erikson, 1996). While in some languages whistling fricatives (stridency) are considered as normal pronunciation (Shosted, 2006), in Serbian (Jovicic et al., 2008), Czech (Honova et al., 2003) as well as some other languages it is considered to be an irregular (pathological) pronunciation.

In the Serbian language, stridence is most commonly manifested in the articulation of fricative /ʃ/ (according to the IPA classification we used /ʃ/ as most similar to Serbian initial fricative in word "ʃuma") as: narrowband stable over time and very intensive resonant occurrence in the diffuse noise spectrum, or a twofold stridence with one stable resonance and one very changeable resonance in the time-frequency representation, or a very short stridence with high variability of the resonant frequency (see Fig. 1). As noted in Jovicic et al. (2008), a strong resonant stridence with an intensity over 20 dB above the envelope of the surrounding spectrum with a minimal duration longer than 10 ms is a necessary condition for resonant occurrence to be perceived as stridence.

The algorithm for automatic stridence detection in the Serbian language (Jovicic et al., 2008) uses Burg's maximum entropy method for calculating stridence measure. This algorithm is accurate in cases where stridence is prominent, with no doubt as to its presence. However, in boundary cases, the algorithm's detection differs from the assessment obtained by the speech therapist. The reason can be found in the fact that the algorithm does not exploit psychoacoustic effects.

In past decades, a variety of computational models of cochlear processing have been developed to provide representations of complex neural activity patterns that arise in the auditory nerve in response to broadband sounds like speech and music (Hohmann, 2002; Patterson and Allerhand, 1995; Patterson and Holdsworth, 1996; Patterson, 2000; Slaney and Lyon, 1993). All of them simulate processing in cochlea using the following: (a) auditory filter-banks which simulate the basilar membrane motion (BMM), (b) some form of compressive adaptation and nonlinearity, for instance a half wave rectifier (HWR) which simulates neural transduction, and (c) temporal integration (Patterson and Allerhand, 1995; Patterson, 2000) or correlogram calculation (Slaney and Lyon, 1993; Slaney et al., 1994) used for generation of the auditory image.

In this paper, a new method for stridence detection based on Paterson's auditory model (Patterson and Allerhand, 1995; Patterson and Holdsworth, 1996; Patterson, 2000) is proposed. Contrary to the original Paterson's auditory model in which strobed integration is applied on each channel independently, we calculated the auditory image along the selected spectral peaks contour. The reason for this is that the central frequency of the signal that represents stridence varies in time across channels of the filter bank.

To optimize tracking of the spectral peaks contour, we modified the nonlinear processing in Paterson's auditory model by replacing the half wave rectifier (HWR) with a calculation of the magnitude of the channel signals. Finally, we simplified Patterson's auditory model by omitting the adaptation in the time domain. This is done because stridence detection is applied to the isolated phonemes whose duration is relatively short compared to the relaxation time of the