Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 59 (2014) 10–21

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Phonetic feature extraction for context-sensitive glottal source processing

John Kane [a,*], Matthew Aylett [b,c], Irena Yanushevskaya [a], Christer Gobl [a]

[a] Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
[b] School of Informatics, University of Edinburgh, UK
[c] CereProc Ltd., UK

Received 19 September 2013; received in revised form 28 November 2013; accepted 24 December 2013
Available online 29 December 2013

Abstract

The effectiveness of glottal source analysis is known to be dependent on the phonetic properties of its concomitant supraglottal features. Phonetic classes like nasals and fricatives are particularly problematic. Their acoustic characteristics, including zeros in the vocal tract spectrum and aperiodic noise, can have a negative effect on glottal inverse filtering, a necessary pre-requisite to glottal source analysis. In this paper, we first describe and evaluate a set of binary feature extractors, for phonetic classes with relevance for glottal source analysis. As voice quality classification is typically achieved using feature data derived by glottal source analysis, we then investigate the effect of removing data from certain detected phonetic regions on the classification accuracy. For the phonetic feature extraction, classification algorithms based on Artificial Neural Networks (ANNs), Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) are compared. Experiments demonstrate that the discriminative classifiers (i.e. ANNs and SVMs) in general give better results compared with the generative learning algorithm (i.e. GMMs). This accuracy generally decreases according to the sparseness of the feature (e.g., accuracy is lower for nasals compared to syllabic regions). We find best classification of voice quality when just using glottal source parameter data derived within detected syllabic regions.

Keywords: Voice quality; Phonation type; Glottal source; Expressive speech; Speech synthesis

1. Introduction

Glottal source analysis refers to the process of trying to parameterise the important and salient aspects of the excitation source for voiced speech, created (mainly) by the vibration of the vocal folds at the larynx. Compared to many other feature extraction methods used in contemporary speech processing, glottal source analysis is relatively complex and involves making several simplifications of the speech production process (for a more comprehensive review of glottal source analysis please refer to: Alku (2011) or Walker and Murphy (2007)). For instance, glottal source analysis typically requires a process known as glottal inverse filtering as a pre-requisite. Glottal inverse filtering is the process of deconvolving a model of the vocal tract transfer function from the speech signal. The process involves making two key (and potentially over-reaching) assumptions.

The first is that speech production can be represented as a Linear Time-Invariant (LTI) system, which facilitates the linear separation of glottal source and vocal tract components (Fant, 1960). This representation is somewhat justified when using short analysis frames, as the articulators in the vocal tract are relatively slowly moving. However, as outlined in several previous publications (see e.g., Lin (1987), Fant and Lin (1987) and Fant et al. (1985b)) source-filter interactions effects exist. These interactions

* Corresponding author. Tel.: +353 1 896 1348.
E-mail addresses: kanejo@tcd.ie (J. Kane), matthewa@cereproc.com (M. Aylett), yanushei@tcd.ie (I. Yanushevskaya), cegobl@tcd.ie (C. Gobl).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.12.003

are most significant in speech regions, for instance, where there is rapid transition of the vocal tract setting within a given analysis frame. The interactions may also be significant when there is a high f_0 and low first formant frequency, as commonly occurs in high vowels. Glottal inverse filtering of such analysis frames may result in an ineffective estimation of the glottal source component.

A second assumption is typically that the vocal tract can be modelled using an all-pole representation. This treatment is usually effective for oral sounds (due to the single-tube characteristic of the vocal tract), but for nasals (i.e. both nasal consonants and nasalised vowels) the different acoustic system is thought to create additional resonances and anti-resonances, and hence pole-zero pairs (Gobl and Mahshie, 2013). The presence of zeros in the vocal tract spectrum may also be true for laterals. As a result, glottal inverse filtering of such regions may be negatively affected by the lack of suitability of the vocal tract all-pole model. Furthermore, it has often been reported that signal processing methods for estimation of the all-pole vocal tract model can be sub-optimal for analysing higher-pitched voices (Alku et al., 2013; Alku and Vilkman, 1994).

One should note that despite these shortcomings for glottal source analysis and criticisms from the literature (notably from Teager and Teager (1990)) the use of glottal source feature data has brought significant benefits to a range of speech technology applications, including: speaker recognition (Chan et al., 2007; Zheng et al., 2007; Murty and Yegnanarayana, 2006), emotion classification (Cullen et al., 2013; Iliev et al., 2010; Lugger and Yang, 2008), characterisation of speaking styles in expressive speech data (Kane et al., 2013a; Székely et al., 2012; Campbell and Mokhtari, 2003), etc. Furthermore, one of the most natural sounding statistical parametric speech synthesisers currently available (Raitio et al., 2011) involves separate modelling of glottal source and vocal tract components, and also allows greater flexibility of voice characteristics compared to conventional methods (Raitio et al., 2014).

However, aside from parametric speech synthesis, which requires modelling of the glottal source for all voiced speech regions, for many other applications (such as those listed above) it may be preferable to use a lesser volume of glottal source feature data but which has been calculated from regions where is most likely to have been derived successfully. Such an approach of deriving glottal source feature data from selective speech regions has previously been suggested (Mokhtari and Campbell, 2003; Mokhtari and Campbell, 2002). Their method involves automatically detecting *centres of reliability*, which they define as vocoids involving high sonorant energy in steady regions where formant estimation is believed to be most reliable. Although they demonstrate the phonetic dependence of a certain glottal source parameter and that this parameter derived in these *centres of reliability* can be effective at discriminating certain affective labels, they do not formally assess the effect of using their selection method compared with not using it.

Recently, we proposed an alternative method for selecting optimal regions for glottal source analysis based on the presence or absence of certain phonetic features (Kane et al., 2013b). In that study we automatically determined the presence of a small number of phonetic features using Mel-Frequency Cepstral Coefficients (MFCCs) as input to Artificial Neural Networks (ANNs). That study revealed that by excluding glottal source feature data in detected nasal and fricative regions significant improvements could be achieved in voice quality classification. Despite these gains, there is still room-for-improvement, in particular in terms of accuracy of the phonetic feature extraction.

Different approaches have been used to automatically derive information on phonetic features from continuous speech. King and Taylor (2000) describe a method based on MFCCs used as inputs to recurrent neural networks and report accuracy in excess of 85% for many features (including vocalic, consonantal, nasal and strident features). However, as the results reported are the % of correct frames (and not, for instance, F-statistics), it is unclear exactly how well the classification performed for sparse features like nasals.

Previous to this, Ali et al. (1999) outlined a system which categorised speech into 4 components (sonorants, stops, fricatives and silences), before further subdividing these into 19 phonetic classes. Experiments on the TIMIT database demonstrated high accuracy, however as before % accuracy is not a very illuminating metric when analysing sparse features. More recently (Tarek and Carson-Berndsen, 2003; Kanokphara et al., 2006), a Hidden Markov Model (HMM) approach to phonetic feature extraction was developed and once more evaluated on the TIMIT database.

Several previous publications have described approaches involving the use of phonetic feature extraction as part of automatic speech recognition systems (Siniscalchi and Lee, 2009; Launay et al., 2002). More recently, authors have looked to exploit the discriminative power of deep neural networks in order to improve phonetic feature extraction accuracy (Siniscalchi et al., 2013; Yu et al., 2012). However, aside from our recent work (Kane et al., 2013b) to the best of our knowledge such approaches have not been investigated in terms of improving glottal source analysis.

1.1. Research questions and aims

The present paper looks to advance the work on phonetic feature extraction by: (1) carrying out a formal evaluation of detection of a range of phonetic features using three different classifiers and (2) by investigating the usefulness of such automatically derived information for glottal source analysis. The research questions can be stated explicitly as: