ARTICLE IN PRESS
JID: YCSLA [m3+sc;October 19, 2016;11:31]

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language xxx (2016) xxx-xxx

www.elsevier.com/locate/csl

Speech vocoding for laboratory phonology

Milos Cernak[a,*], Stefan Benus[b], Alexandros Lazaridis[a]

[a] *Idiap Research Institute, Martigny, Switzerland*

[b] *Constantine the Philosopher University in Nitra, Slovakia and Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia*

Received 9 February 2015; received in revised form 15 September 2016; accepted 5 October 2016

Abstract

Using phonological speech vocoding, we propose a platform for exploring relations between phonology and speech processing, and in broader terms, for exploring relations between the abstract and physical structures of a speech signal. Our goal is to make a step towards bridging phonology and speech processing and to contribute to the program of Laboratory Phonology.

We show three application examples for laboratory phonology: compositional phonological speech modelling, a comparison of phonological systems and an experimental phonological parametric text-to-speech (TTS) system. The featural representations of the following three phonological systems are considered in this work: (i) Government Phonology (GP), (ii) the Sound Pattern of English (SPE), and (iii) the extended SPE (eSPE). Comparing GP- and eSPE-based vocoded speech, we conclude that the latter achieves slightly better results than the former. However, GP − the most compact phonological speech representation − performs comparably to the systems with a higher number of phonological features. The parametric TTS based on phonological speech representation, and trained from an unlabelled audiobook in an unsupervised manner, achieves intelligibility of 85% of the state-of-the-art parametric speech synthesis.

We envision that the presented approach paves the way for researchers in both fields to form meaningful hypotheses that are explicitly testable using the concepts developed and exemplified in this paper. On the one hand, laboratory phonologists might test the applied concepts of their theoretical models, and on the other hand, the speech processing community may utilize the concepts developed for the theoretical phonological models for improvements of the current state-of-the-art applications.

Keywords: Phonological speech representation; Parametric speech synthesis; Laboratory phonology

1. Introduction

Speech is a domain exemplifying the dichotomy between the continuous and discrete aspects of human behaviour. On the one hand, the articulatory activity and the resulting acoustic speech signal are continuously varying. On the other hand, for speech communication to convey meaning, this continuous signal must be, at the same time, perceivable as contrastive. Traditionally, these two aspects have been studied within phonetics and phonology respectively. Following significant successes of this dichotomous approach, for example in speech synthesis and recognition, recent decades have witnessed a lot of progress in understanding and formal modelling of the relationship between these two aspects, e.g. the program of Laboratory Phonology (Pierrehumbert et al., 2000) or the

* Corresponding author
E-mail addresses: milos.cernak@idiap.ch, milos_cernak@hotmail.com (M. Cernak).

http://dx.doi.org/10.1016/j.csl.2016.10.001
0885-2308/ 2016 Published by Elsevier Ltd.

renewed interest in the approaches based on Analysis by Synthesis (Bever and Poeppel, 2010; Hirst, 2011). The goal of this paper is to follow these developments by proposing a platform for exploring relations between the mental (abstract) and physical structures of the speech signal. In this, we aim at mutual cross-fertilisation between phonology, as a quest for understanding and modelling of cognitive abilities that underlie systematic patterns in our speech, and speech processing, as a quest for natural, robust, and reliable automatic systems for synthesising and recognising speech.

As a first step in this direction we examine a cascaded speech analysis and synthesis approach (known also as vocoding) based on phonological representations and how this might inform both quests mentioned above. In parametric vocoding speech segments of different time-domain granularity, ranging from speech frames, e.g. in the formant (Holmes, 1973), or articulatory (Goodyear and Wei, 1996; Laprie et al., 2013) domains, to phones (Lee and Cox, 2001; Tokuda et al., 1998), and syllables (Cernocky et al., 1998), are used in sequential processing. In addition to these segments, phonological representations have also been shown to be useful for speech processing e.g. by King and Taylor (2000). In our work, we explore a direct link between phonological features and their engineered acoustic realizations. In other words, we believe that abstract phonological sub-segmental, segmental, and suprasegmental structures may be related to the physical speech signal through a speech engineering approach, and that this relationship is informative for both phonology and speech processing.

The motivation for this approach is two-fold. Firstly, phonological representations (together with grammar) create a formal model whose overall goal is to capture the core properties of the cognitive system underlying speech production and perception. This model, linking subsegmental, segmental, and suprasegmental phonological features of speech, finds independent support in the correspondence between a) the brain-generated cortical oscillations in the 'delta' (1−3 Hz), 'theta' (4−7 Hz), and faster 'gamma' ranges (25−40 Hz), and b) the temporal scales for the domains of prosodic phrases, syllables, and certain phonetic features respectively. In this sense, we may consider phonological representations embodied (Giraud and Poeppel, 2012). Hence, speech processing utilizing such a system might lead to a biologically sensible and empirically testable computational model of speech.

Secondly, phonological representations are inherently multilingual (Siniscalchi et al., 2012). This in turn has an attractive advantage in the context of multilingual speech processing in lessening the reliance on purely phonetic decisions. The independence of the phonological representations from a particular language on the one hand and the availability of language specific mapping between these representations and the acoustic signal through speech processing methods on the other hand, offer (we hope) a path towards a context-based interpretation of the phonological representation that is grounded in phonetic substance but at the same time abstract enough to allow for a more streamlined approach to multilingual speech processing.

In this work, we propose to use the phonological vocoding of Cernak et al. (2015) and other advances of speech processing for testing certain aspects of phonological theories. We consider the following phonological systems in this work:

- The Government Phonology (GP) features (Harris and Lindsey, 1995) describing sounds by fusing and splitting of 11 primes.
- The Sound Pattern of English (SPE) system with 13 features established from natural (articulatory) features (Chomsky and Halle, 1968).
- The extended SPE system (eSPE) (Siniscalchi et al., 2012; Yu et al., 2012) consisting of 21 phonological features.

Having trained phonological vocoders for the three phonological models of sound representation, we describe several application examples combining speech processing techniques and phonological representations. Our primary goal is to demonstrate the usefulness of the analysis by synthesis approach by showing that (i) the vocoder can generate acoustic realizations of phonological features used by compositional speech modelling, (ii) speech sounds (both individual sounds not seen in training and intelligible continuous speech) can be generated from the phonological speech representation, and (iii) the testing of hypotheses relating phonetics and phonology is possible; we test the hypothesis that the best phonological speech representation achieves the best quality vocoded speech, by evaluating the phonological features in both directions, recognition and synthesis, simultaneously. Additionally, we compare the segmental properties of the phonological

Please cite this article as: M. Cernak et al., Speech vocoding for laboratory phonology, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.10.001