Speech Communication 84 (2016) 36–45

Contents lists available at ScienceDirect

Speech Communication

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

On structured sparsity of phonological posteriors for linguistic parsing

Milos Cernak[a,*], Afsaneh Asaei[a], Hervé Bourlard[a,b]

[a] Idiap Research Institute, Martigny, Switzerland
[b] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ARTICLE INFO

Article history:
Available online 3 September 2016

Keywords:
Phonological posteriors
Structured sparse representation
Deep neural network (DNN)
Binary pattern matching
Linguistic parsing

ABSTRACT

The speech signal conveys information on different time scales from short (20–40 ms) time scale or segmental, associated to phonological and phonetic information to long (150–250 ms) time scale or supra segmental, associated to syllabic and prosodic information. Linguistic and neurocognitive studies recognize the *phonological* classes at segmental level as the essential and invariant representations used in speech temporal organization.

In the context of speech processing, a deep neural network (DNN) is an effective computational method to infer the probability of individual phonological classes from a short segment of speech signal. A vector of all phonological class probabilities is referred to as *phonological posterior*. There are only very few classes comprising a short term speech signal; hence, the phonological posterior is a sparse vector. Although the phonological posteriors are estimated at segmental level, we claim that they convey supra-segmental information. Specifically, we demonstrate that phonological posteriors are indicative of syllabic and prosodic events.

Building on findings from converging linguistic evidence on the gestural model of Articulatory Phonology as well as the neural basis of speech perception, we hypothesize that phonological posteriors convey properties of linguistic classes at multiple time scales, and this information is embedded in their support (index) of active coefficients. To verify this hypothesis, we obtain a binary representation of phonological posteriors at the segmental level which is referred to as first-order sparsity structure; the high-order structures are obtained by the concatenation of first-order binary vectors. It is then confirmed that the classification of supra-segmental linguistic events, the problem known as *linguistic parsing*, can be achieved with high accuracy using a simple binary pattern matching of first-order or high-order structures.

1. Introduction

A theory of Articulatory Phonology (Browman and Goldstein, 1986) suggests that an utterance is described by temporally overlapped distinctive constriction actions of the vocal tract organs, known as gestures. Gestures are changes in the vocal tract, such as opening and closing, widening and narrowing, and they are phonetic in nature (Fowler et al., 2015). Gestures compose units of information and can be used to distinguish words in all languages. Recent work on Articulatory Phonology (Goldstein and Fowler, 2003) further suggests an existence of coupling/synchronisation of gestures that influence the syllable structure of an utterance.

Phonological classes (e.g., (Chomsky and Halle, 1968; Jakobson and Halle, 1956)) emerge during the phonological encoding process – the processes of speech planning for articulation, namely the preparation of an abstract phonological code and its transformation into speech motor plans that guide articulation (Levelt, 1993). Stevens (2005) reviews evidence about a universal set of phonological classes that consists of articulator-bound classes and articulator-free classes ([continuant], [sonorant], [strident]). We support the hypothesis in this work and consider phonological classes in our work as essential and invariant acoustic-phonetic elements used in both linguistics and cognitive neuroscience studies for speech temporal organization.

In the present paper, we study inferred phonological posterior features that consist of phonological class probabilities given a segment of input speech signal. The class-conditional posterior probabilities are estimated using a Deep Neural Network (DNN). Cernak et al. (2015b) introduce the phonological posterior fea-

* Corresponding authors. both authors contributed equally to this manuscript.
E-mail addresses: milos.cernak@idiap.ch (M. Cernak), afsaneh.asaei@idiap.ch (A. Asaei), herve.bourlard@idiap.ch (H. Bourlard).

http://dx.doi.org/10.1016/j.specom.2016.08.004
0167-6393/

tures for phonological analysis and synthesis, and we hypothesise their relation to the linguistic gestural model. Saltzman and Munhall (1989) describe the constriction dynamics model as a computational system that incorporates the theory of articulatory phonology. This gestural model defines gestural scores as the temporal activation of each gesture in an utterance. Thus, we hypothesise that phonological posteriors are related to gestural scores and that the trajectories of phonological posteriors correspond to the distal representation of articulatory gestures. In a broader view, we consider the trajectories of phonological posteriors as articulatory-bound and articulatory-free gestures. Since gestures are linguistically relevant (Liberman and Whalen, 2000), we hypothesize that phonological posteriors should convey supra-segmental information through their inter-dependency low-dimensional structures. Hence, by characterizing the structure of phonological posteriors, it should be possible to perform a top-down linguistic parsing, i.e., by knowing *a priori* where linguistic boundaries lie.

Previously in (Asaei et al., 2015), we have shown that phonological posteriors admit sparsity structures underlying short-term segmental representations where the structures are quantified as sparse binary vectors. In this work, we explore this idea further and consider trajectories of phonological posteriors for supra-segmental structures. We show that unique structures (codes) exist for distinct linguistic classes and identification of these structures enables us to perform linguistic parsing. The linguistic parsing is thus achieved through identification of low dimensional sparsity structures of phonological posteriors followed by binary pattern matching. This idea is in line with an assumption that physical and cognitive speech structures are, in fact, the low and high dimensional descriptions of a single (complex) system[1].

Our contribution to advance the study of phonological posteriors is two-fold: First, we review converging evidence from linguistic and neural basis of speech perception, that support the hypothesis about phonological posteriors conveying properties of linguistic classes at multiple time scales. Second, we propose linguistic parsing based on structured sparsity as low dimensional characterization of phonological posteriors.

The rest of the paper is organized as follows. Section 2 provides review of the definition and relation of phonological posteriors to the linguistic gestural model and subsequently to cognitive neuroscience, Section 3 introduces linguistic parsing, and Section 4 presents the details of experimental analysis. Finally, Section 5 concludes the paper and discusses the results in a broader context.

2. Phonological class-conditional posteriors

Fig. 1 illustrates the process of the phonological analysis (Cernak et al., 2015b; Yu et al., 2012). The phonological posterior features are extracted starting with converting a segment of speech samples into a sequence of acoustic features $X = \{\vec{x}_1, \ldots, \vec{x}_n, \ldots, \vec{x}_N\}$ where N denotes the number of segments in the utterance. Conventional cepstral coefficients can be used as acoustic features. Then, a bank of phonological class analysers realised via neural network classifiers converts the acoustic feature observation sequence X into a sequence of phonological posterior probabilities $Z = \{\vec{z}_1, \ldots, \vec{z}_n, \ldots, \vec{z}_N\}$; a posterior probability $\vec{z}_n = [p(c_1|x_n), \ldots, p(c_k|x_n), \ldots, p(c_K|x_n)]^\top$ consists of K phonological class-conditional posterior probabilities where c_k denotes the phonological class and $.^\top$ stands for the transpose operator.

The phonological posteriors Z yield a parametric speech representation, and we hypothesise that the trajectories of the articulatory-bound phonological posteriors correspond to the distal representation of the gestures in the gestural model of speech production (and perception). For example, Fig. 2 shows a comparison of articulatory tongue tip gestures (vertical direction with respect to the occlusal plane) and the phonological anterior posterior features, on an electromagnetic articulography (EMA) recording (Lee et al., 2005). The articulatory gesture and phonological posteriors trajectory have the same number of maximums, and their relation is evident. A more asynchronous relation toward the end of utterance is caused by asynchronous relations of fast tongue tip movements causing an obstruction in the mouth and generated acoustics.

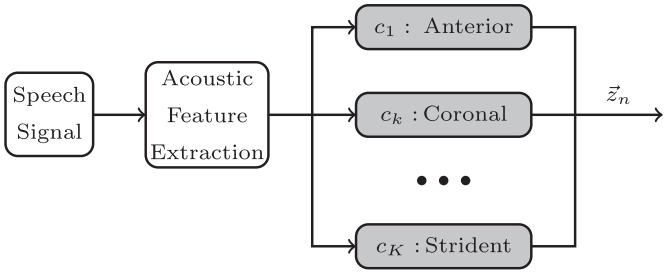

Fig. 1. The process of phonological analysis. Each segment of speech signal is represented by phonological posterior probabilities $\vec{z}_n$ that consist of K class-conditional posterior probabilities. For each phonological class, a DNN is trained to estimate its posterior probability given the input acoustic features.

The hypothesis of correspondence of the phonological posterior features to the gestural trajectories is also motivated by the analogy to the constriction dynamics model (Saltzman and Munhall, 1989) that takes gestural scores as input and generates articulator trajectories and acoustic output. Alternatively to this constriction dynamics model, we generate acoustic output using a phonological synthesis DNN described in Cernak et al. (2015b).

In the following sections, we outline converging evidence from linguistics as well as the neural basis of speech perception, that support the hypothesis about phonological posteriors conveying properties of linguistic classes at multiple time scales.

2.1. Linguistic evidence

Linguistics defines two traditional components of speech structures:

1. Cognitive structure consisting of system primitives, that is, the units of representation for cognitively relevant objects such as phonemes or syllables. The system primitives are represented by *canonical* phonological features (classes) that emerge during the phonological encoding process (Levelt, 1993).
2. Physical structure generated by a set of permissible operations over cognitive system primitives that yield the observed (surface) patterns. The physical structure is represented by *surface* phonological features, continuous variables that may be partially estimated from the speech signal by inverse filtering. Phonological posteriors can also be classified as surface phonological features.

The canonical (discrete) phonological features have been used over the last 60 years to describe cognitive structures of speech sounds. Miller and Nicely (1955) have experimentally shown that consonant confusions were perceived similarly for the isolated phonemes and the phonemes within the perceptual groups they form (for example /t/, /p/ and /k/ form a perceptual group, characterised by binary features of voiceless stops). Canonical features are extensively studied in phonology. In the tradition of Jakobson and Halle (1956) and Chomsky and Halle (1968), phonemes are assumed to consist of feature bundles – the Sound Pattern of English (SPE). Later advanced phonological systems were proposed,

[1] http://www.haskins.yale.edu/research/gestural.html.