IDIAP RESEARCH REPORT

RESEARCH INSTITUTE

ON STRUCTURED SPARSITY OF PHONOLOGICAL POSTERIORS FOR LINGUISTIC PARSING

Milos Cernak   Afsaneh Asaei   Hervé Bourlard

Idiap-RR-07-2016

APRIL 2016

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

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*

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. Namely, 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 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 segmental level which is referred to as first-order sparsity structure; the high-order structures are obtained by concatenation of first-order binary vectors. It is then confirmed that 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.

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

*Corresponding authors; both authors contributed equally to this manuscript.

Email addresses: `milos.cernak@idiap.ch` (Milos Cernak), `afsaneh.asaei@idiap.ch` (Afsaneh Asaei), `herve.bourlard@idiap.ch` (Hervé Bourlard)

1. Introduction

A theory of Articulatory Phonology (Browman and Goldstein, 1986) suggests that an utterance is described by temporally overlapped (co-articulated) distinctive constriction actions of the vocal tract organs, actions 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., (Jakobson and Halle, 1956; Chomsky and Halle, 1968)) 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 follow the Stevens's view 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 features 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 computational system that incorporates Articulatory Phonology approach. This gestural model defines gestural scores as the temporal activation of each gesture in an utterance. Thus, we hypothesise relation of the gestural scores to phonological posteriors, 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) exists 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 about 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 cross-field context.

2. Phonological Class-conditional Posteriors

Figure 1 shows a process of the phonological analysis (Yu et al., 2012; Cernak et al., 2015b). The phonological posterior features are extracted by phonological analysis that starts by 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.

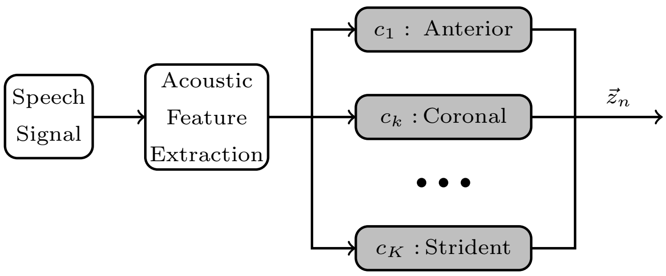


Figure 1: *The process of phonological analysis. Each segment of speech signal is represented by phonological posterior probabilities $\vec{z}_n$ that consist of K class-conditional posterior probabilities. For each phonological class, a DNN is trained to estimate its posterior probability given the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

The 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

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

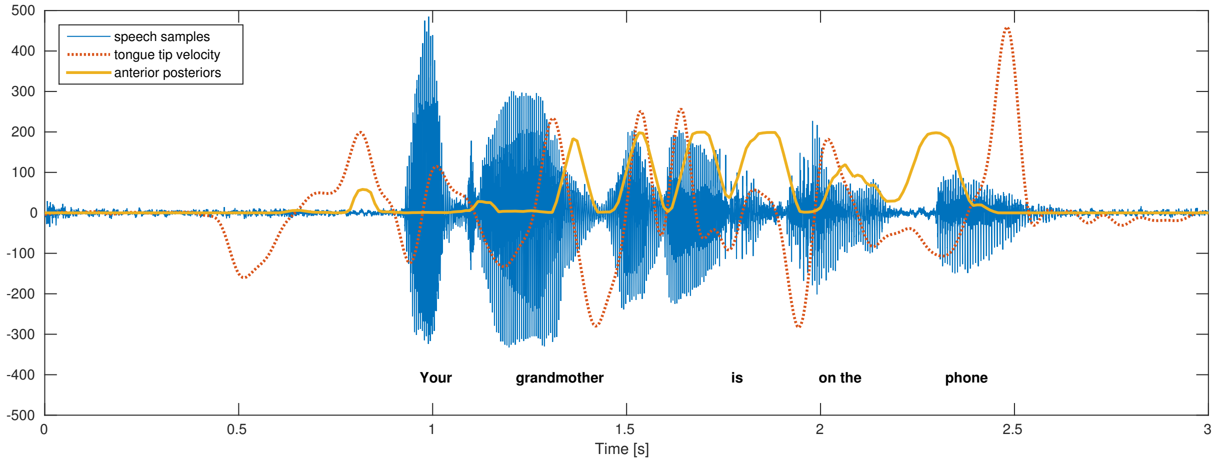


Figure 2: *Anterior phonological posteriors vs. the electromagnetic articulography tongue tip measurement. The correlation between the articulatory gesture score and the trajectory of its corresponding phonological class posterior probability is evident.*

gestures in the gestural model of speech production (and perception). For example, Figure 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 EMA recording (Lee et al., 2005). The articulatory gesture and phonological posteriors trajectory have the same number of maximums, and their relation is evident.

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

In the following sections, we outline converging evidence from linguistics as well as 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* phonologi-

cal features, continuous variables that may be partially estimated from the speech signal by inverse filtering. Phonological posteriors can be also 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 from the observed and binary confusion values (a consonant present or not in a group of consonants). 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, such as multi-valued phonological features of Ladefoged and Johnson (2014), and monovalent Government Phonology features of Harris and Lindsey (1995) that describe sounds by fusing and splitting of primes.

The surface code includes co-articulated canonical code, with further intrinsic (speaker-based) and extrinsic (channel-based) speech variabilities that contribute to the opacity of the function operating between the two codes. The surface features may contain additional gestures dependent on the prosodic context, such as position within a syllable, word, and sentence. Other changes in surface phonological features at different time granularities are due to phonotactic constraints. For example, glides are always syllable-initial, and consonants that follow a non-tense vowel are always in the coda of the syllable (Stevens, 2005).

Browman and Goldstein (1986, 1989, 1992) introduced articulatory gestures as basis for human speech production. The trajectories of gestures contain overlapping, interleaving, and merging segments, as a result of co-articulation. It is said that gestures are phonetic in nature (Fowler et al., 2015). Browman and Goldstein (1988); Nam et al. (2009) provide direct evidence on existence of supra-segmental (syllable) structures in gestures. For example, the task dynamic model of inter-articulator coordination in speech (Saltzman and Munhall, 1989) implements a syllable structure-based gesture coupling model.

2.2. Cognitive Neuroscience Evidence

Modern cognitive neuroscience studies use phonological classes as essential and invariant acoustic-phonetic primitives for speech temporal organization (Poeppel, 2014). Neurological data from the brain activity during speech planning, production or perception are increasingly used to inform such cognitive models of speech and language.

The auditory pre-processing is done in the cochlea, and then split into two parallel pathways leading from the auditory system (Wernicke, 1874/1969). For example, the dual-stream model of the functional anatomy of language (Hickok and Poeppel, 2007) consists of a ventral stream: sound to meaning function using phonological classes, phonological-level processing at superior temporal sulcus bilaterally, and a dorsal stream: sound to action, a direct link between sensory and motor representations of speech based again on the phonological classes. The former stream supports the speech perception, and the latter stream reflects the observed disruptive effects of altered auditory feedback on speech production. Phillips et al. (2000); Mesgarani et al. (2014) present evidence of discrete phonological classes available in the human auditory cortex.

Recent evidence from psychoacoustics and neuroimaging studies indicate that auditory cortex segregates information emerging from the cochlea on at least three discrete time-scales processed in the auditory cortical hierarchy: (1) "stress" δ frequency (1–3 Hz), (2) "syllabic" θ frequency (4–8 Hz) and (3) "phonetic" low γ frequency (25–35 Hz) (Giraud and Poeppel, 2012). Leong et al. (2014) show that phase relations between the phonetic and syllabic amplitude modulations, known as hierarchical phase locking and nesting or synchronization across different temporal granularity (Lakatos et al., 2005), is a good indication of the syllable stress. Intelligible speech representation with stress and accent information can be constructed by asynchronous fusion of phonetic and syllabic information (Cernak et al., 2015a).

In addition, not only phase locking across different temporal granularity has linguistic interpretation. Bouchard et al. (2013) claim that functional organisation of ventral sensorimotor cortex supports gestural model developed in Articulatory Phonology. Analysis of spatial patterns of activity showed a hierarchy of network states that organizes phonemes by articulatory-bound phonological features. Leonard et al. (2015) further show how listeners use phonotactic knowledge (phoneme sequence statistics) to process spoken input and to link low-level acoustic representations (the coarticulatory dynamics of the sounds through the encoding of combination of phonological features) with linguistic information about word identity and meaning. This is converging evidence on the relation of linguistic gestural model and speech and language cognitive neuroscience models on phonological class-conditional posteriors used in our work.

3. Sparse Phonological Structures for Linguistic Parsing

Building on linguistic and cognitive findings, the phonological representation of speech lies at the center of human speech processing. Speech analysis is performed at different time granularity broadly categorized as segmental and supra-segmental levels. The phonological classes define the sub-phonetic and phonetic attributes recognized at segmental level whereas the syllables, lexical stress and prosodic accent are the basic supra-segmental events - c.f. Figure 3. The phonological representations are often studied at segmental level and their supra-segmental properties are not investigated. It is this supra-segmental characterization of phonological posteriors that this manuscript will explore.

3.1. Structured Sparsity of Phonological Posteriors

Phonological posteriors are indicators of the physiological posture of human articulation machinery. Due to the physical constraints, only few combinations can be realized in our vocalization. This physical

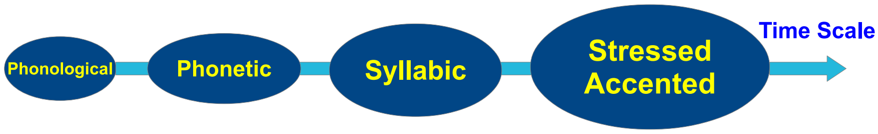


Figure 3: *Different time granularity of speech processing. The phonological and phonetic classes are segmental attributes whereas the syllable type, stress and accent are linguistic events recognized at supra-segmental level. Inferring the supra-segmental attributes from sub-phonetic features is the task of linguistic parsing (Poeppel, 2003).*

limitation leads to a small number of unique patterns exhibited over the entire speech corpora (Asaei et al., 2015). We refer to this structure as *first-order structure* which is exhibited at segmental level.

Moreover, the dynamics of the structured sparsity patterns is slower than the short segments and it is indicative of supra-segmental information, leading to a higher order structure underlying a sequence (trajectory) of phonological posteriors. This structure is exhibited at supra-segmental level by analyzing a long duration of phonological posteriors, and it is associated to the syllabic information or more abstract linguistic attributes. We refer to this structure as *high-order structure*.

We hypothesize that the first-order and high-order structures underlying phonological posteriors can be exploited as indicators of supra-segmental linguistic events. To test this hypothesis, we identify all structures exhibited in different linguistic classes. The set of class-specific structures is referred to as the codebook.

3.2. Codebook of Linguistic Structures

The goal of codebook construction is to collect all the structures associated to a particular linguistic event. To that end, we consider *binary* phonological posteriors where the probabilities above 0.5 are normalized to 1 and the probabilities less than 0.5 are forced to zero. This rounding procedure enables us to identify the active phonological components as indicators of linguistic events. It also alleviates the speaker and environmental variability encoded in the continuous probabilities. An immediate extension to this approach is multi-valued quantization of phonological posteriors as opposed to 1-bit quantization. We consider this extension for our future studies and focus on binary phonological indicators to obtain linguistic structures.

Different codebooks are constructed for different classes. Namely, one codebook encapsulates all the binary structures of the consonants whereas another codebook has all the binary structures of the vowels. These two codebooks will be used for binary pattern matching to classify consonants versus vowels as will be explained in the next Section 3.3. Likewise, one codebook encapsulates all the binary structures of stressed syllables whereas another codebook has all the binary structures of unstressed syllables, and these two codebooks are used for stress detection; the similar procedure holds for accent detection.

The codebook can be constructed from the first-order structures as well as the high-order structures. For example, a second-order codebook is formed from all the binary structures of second-order phonological posteriors obtained by concatenation of two adjacent phonological posteriors to form a super vector from the segmental representations.

The procedure of codebook construction for classification of linguistic events rely on the assumption that there are unique structures per class (consonant, stressed or syllable) and the number of permissible patterns is small. Hence, classification of any phonological posterior can be performed by finding the closest match to its binary structure from the codebooks characterizing different linguistic classes.

3.3. Pattern Matching for Linguistic Parsing

Figure 3 illustrates different time granularity identified for processing of speech. Inferring the supra-segmental properties such as syllable type or accented / stressed pronunciation is known as linguistic parsing (Poeppel, 2003). Parsing can be performed in a top-down procedure, driven by a-priori known segment boundaries.

Having the codebooks of structures underlying phonological posteriors, linguistic parsing amounts to binary pattern matching. The similarity metric plays a critical role in classification accuracy. Hence, we investigate several metrics found effective in different binary classification settings. The definition of binary similarity measures are expressed by *operational taxonomic units* (Dunn and Everitt, 1982). Consider two binary vectors i, j: a denotes the number of elements where the values of both i, j are 1, meaning *"positive match"*; b denotes the number of elements where the values of i, j is $(0,1)$, meaning *"i absence mismatch"*; c denotes the number of elements where the values of i, j is $(1,0)$, meaning *"j absence mismatch"*; d denotes the number of elements where the values of both i, j are 0, meaning *"negative match"*. The definition of binary similarity measures used for our evaluation of linguistic parsing is as follows (Choi and Cha, 2010):

$$S_{\text{JACCARD}} = \frac{a}{a+b+c} \tag{1}$$

$$S_{\text{INNERPRODUCT}} = a + d \tag{2}$$

$$S_{\text{HAMMING}} = b + c \tag{3}$$

$$S_{\text{AMPLE}} = \frac{a(c+d)}{c(a+b)} \tag{4}$$

$$S_{\text{SIMPSON}} = \frac{a}{\min(a+b, a+c)} \tag{5}$$

$$S_{\text{HELLINGER}} = 2\sqrt{1 - \frac{a}{\sqrt{(a+b)(a+c)}}} \tag{6}$$

Different metrics are motivated due to different treatment of positive/negative match and mismatches in indicators of phonological classes. The most effective similarity measure for linguistic parsing can imply different cognitive mechanism governing human perception of linguistic attributes.

In the top-down approach to linguistic parsing, syllable boundaries are first estimated from the speech signal. Then, the similarity between the class-specific codebook members and a phonological posterior is measured. The class label is determined based on the maximum similarity. We provide empirical results on linguistic parsing in the following Section 4.

4. Experiments

4.1. Experimental setup

We use an open-source phonological vocoding platform[2] to obtain phonological posteriors. Briefly, the platform is based on cascaded speech analysis and synthesis that works internally with the phonological speech representation. In the phonological analysis part, phonological posteriors are detected directly from the speech signal by a bank of parallel Deep Neural Networks (DNNs). Each DNN determines the probability of a particular phonological class. In the following, we describe the databases and DNN training procedure to estimate the phonological posterior features.

4.1.1. Speech Databases

To confirm that uniqueness of class-specific sparsity structures is a language-independent property, we conducted our evaluations on English and French speech corpora. Accordingly, to confirm independence of the proposed methodology on a phonological system, two different phonological speech representations are considered: the SPE feature set (Chomsky and Halle, 1968), and the extended SPE feature set (Cernak et al., 2015b) are used in training of the DNNs for phonological posterior estimation on English and French data respectively. Table 1 lists data used in the experimental setup.

Table 1: *Data used for DNN training to obtain phonological posteriors, and evaluation data.*

Purpose	Database	Size (hours)
Training English data	WSJ	66
Training French data	Ester	58
Evaluation English data	Nancy	1.5
Evaluation French data	SIWIS	1

To train the DNNs for phonological posterior estimation on English data, we use the Wall Street Journal (WSJ0 and WSJ1) continuous speech recognition corpora (Paul and Baker, 1992). To train the DNNs for phonological posterior estimation on French data, we use the Ester database (Galliano et al., 2006) containing standard French radio broadcast news in various recording conditions.

[2] `https://github.com/idiap/phonvoc`

Once DNNs are trained, the phonological posterior features are estimated for the Nancy and SIWIS recordings which is used for the subsequent cross-database linguistic parsing experiments.

The Nancy database is provided in Blizzard Challenge[3]. The speaker is known as "Nancy", and she is a US English native female speaker. The database consists of 16.6 hours of high quality recordings of natural expressive human speech made in an anechoic chamber at a 96K sampling rate during 2007 and 2008. The audio of the last 1.5 hours of the recordings was selected and re-sampled to sampling frequency of 16kHz for our experiments. The transcription of the audio data comprised of around 12k utterances. The text was processed by a conventional and freely available TTS front-end (Black et al., 1997), resulting the segmental (quinphone phonetic context) and supra-segmental (full-context) labels. The full-context labels included binary lexical stress and prosodic accents. The labels were forced aligned with the audio recordings.

The SIWIS database[4] consists of 26 native French speakers. The labels were obtained using forced alignment. We generated full-context labels using the French text analyzer eLite (Roekhaut et al., 2014). Unlike Nancy speech recordings, SIWIS data is noisy and recorded in less restricted acoustic conditions. Evaluations on both English and French corpora enables us to confirm and compare the applicability of our linguistic parsing method across languages with different phonological classes as well as uncontrolled recording scenarios.

4.1.2. DNN Training for Phonological Posterior Estimation

First, we trained a phoneme-based automatic speech recognition system using mel frequency cepstral coefficients (MFCC) as acoustic features. The phoneme set comprising of 40 phonemes (including "sil", representing silence) was defined by the CMU pronunciation dictionary. The three-state, cross-word triphone models were trained with the HTS variant (Zen et al., 2007) of the HTK toolkit on the 90% subset of the *si_tr_s_284* set. The remaining 10% subset was used for cross-validation. The acoustic models were used to get boundaries of the phoneme labels.

Then, the labels of phonemes were mapped to the SPE phonological classes. In total, K DNNs were trained as the phonological analyzers using the short segment (frame) alignment with two output labels indicating whether the k-th phonological class exists for the aligned phoneme or not. The number of K is determined from the set of phonological classes and it is equal to 15 for the English data, and 24 for the French data. The DNNs have the architecture of 351x1024x1024x1024x2 neurons, determined empirically. The input vectors are 39 order MFCC features with the temporal context of 9 successive frames. The parameters were initialized using deep belief network pre-training done by single-step contrastive divergence (CD-1) procedure of Hinton et al. (2006). The DNNs with the softmax output function were then trained using a mini-batch based stochastic gradient descent algorithm with the cross-entropy cost function of the

[3]`http://www.cstr.ed.ac.uk/projects/blizzard/2011/lessac_blizzard2011`

[4]`https://www.idiap.ch/project/siwis/downloads/siwis-database`

KALDI toolkit (Povey et al., 2011). Table 2 lists the detection accuracy for different phonological classes. The DNNs outputs for individual phonological classes determine the phonological posterior probabilities.

Similarly, we trained French phonological posterior estimators. The phoneme set comprising 38 phonemes (including “sil”) was defined by the BDLex (Perennou, 1986) lexicon. The aligned phoneme labels were mapped to the French extended SPE (eSPE) phonological classes. The DNN architecture is similar to the English data, and it is initialized by deep belief network pre-training. Table 3 lists the detection accuracy for various eSPE classes.

Table 2: *Classification accuracies (%) of English phonological class detectors on train and cross-validation (CV) data.*

Phonological Classes	Accuracy (%)		Phonological Classes	Accuracy (%)	
	Train	CV		Train	CV
vocalic	97.3	96.5	round	98.7	98.1
consonantal	96.3	95.0	tense	96.6	95.3
high	97.0	95.7	voice	96.5	95.6
back	96.2	94.8	continuant	97.3	96.3
low	98.4	97.6	nasal	98.9	98.4
anterior	96.8	95.6	strident	98.7	98.2
coronal	96.1	94.6	rising	98.6	97.8

4.2. Linguistic Parsing

In this section, we present the evaluation results of our proposed method of top-down linguistic parsing. We provide empirical results on sparsity of phonological posteriors and confirm validity of class-specific codebooks to classify supra-segmental linguistic events based on binary pattern matching.

4.2.1. Binary Sparsity of Phonological Posteriors

Figure 4 illustrates a histogram of phonological posteriors distribution. We can see that the distribution exhibits the binary nature of phonological posterior being valued in the range of $[0-1]$, and mostly concentrated very close to either 1 or 0. This binary pattern is visible for both stressed and unstressed syllables as demonstrated in the right and left plots, respectively.

The 1-bit discretization, achieved by rounding of posteriors results into a very small number of unique phonological binary structures, counting merely 0.1% of all possible structures. This imply that the binary patterns may encode particular shapes of the vocal tract. Since a limited number of these shapes can be created for human speech, the number of unique patterns is very small.

Table 3: *Classification accuracies (%) of the French phonological class detectors on train and cross-validation (CV) data.*

Phonological Classes	Accuracy (%) Train	CV	Phonological Classes	Accuracy (%) Train	CV
Labial	98.2	97.4	Nasal	99.0	98.8
Dorsal	97.3	96.3	Stop	97.6	97.0
Coronal	95.9	94.7	Approximant	98.2	97.6
Alveolar	98.9	98.4	Anterior	95.4	94.2
Postalveolar	99.7	99.5	Back	98.0	97.1
High	97.0	95.9	Lennis	98.0	97.4
Low	97.4	96.5	Fortis	97.5	96.8
Mid	96.9	96.2	Round	97.3	96.6
Uvular	98.7	98.1	Unround	95.9	95.1
Velar	99.2	98.8	Voiced	95.4	94.3
Vowel	94.3	93.1	Central	98.5	98.1
Fricative	97.1	96.1	Silence	97.8	97.4

This property encouraged us also to use this binary approximation in low bit-rate speech coding (Cernak et al., 2015b; Asaei et al., 2015); these studies confirmed that binary approximation has only a negligible impact on perceptual speech quality.

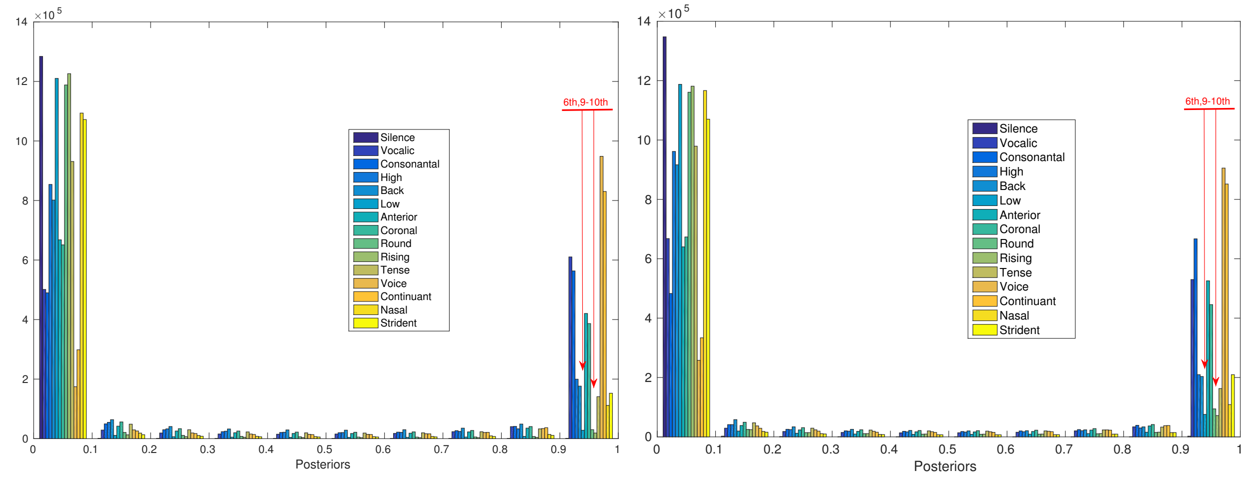
Furthermore, comparing Figures 4a and 4b, we can observe that at least the [low] (6th), [round] (9th) and [rising] (10th) classes are significantly more present in stressed binary-ones than in unstressed syllables. This observation indicates that stressed syllables are more prominent in prosodic typology (e.g., (Jun, 2005)) – mouth are more open. We use the [rising] feature to differentiate diphthongs from monophthongs, that is also more prominent in stressed syllables.

4.2.2. Class-specific Linguistic Structures

The objective of this section is to confirm the hypothesis that phonological posteriors admit class-specific structures which can be used for identification of supra-segmental linguistic events.

Following the procedure of codebook construction elaborated in Section 3.2, we obtain six different codebooks to address the following parsing scenarios:

- Consonant vs. vowel (C-V) detection.
- Stress vs. unstressed detection.
- Accented vs. unaccented detection.



(a) Unstressed Syllables

(b) Stressed Syllables

Figure 4: *Binary sparsity of continuous phonological posteriors Z. We can observe that at least the [low] (6th), [round] (9th) and [rising] (10th) classes are significantly more present in stressed binary-ones than in unstressed syllables.*

The size of codebooks equals to the number of unique binary class-specific structures. The number of unique structures is indeed a small fraction of the whole speech data. For example, the ratio of unique binary structures for the whole Nancy database (16.6 hours of speech) is about 0.08% of the total number of phonological posteriors.

The detection method relies on binary pattern matching and the codebook with a member which possesses maximum similarity to the phonological posterior determines its supra-segmental linguistic property, i.e. being a consonant or vowel, stressed or unstressed and accented or unaccented. The three parsing scenarios are tested separately so the linguistic parsing amounts to a binary classification problem.

We process each speech segments independently. To obtain a decision for the supra-segmental events from the segmental labels, the labels of all the segments comprising a supra-segmental event are pulled to form a decision based on majority counting. In other words, the number of segments being recognized as a particular event is counted, and the final supra-segmental label is decided according to the maximum count. If the similarities of a binary phonological posterior to both codebooks are equal, the segment is not labeled, thus excluded from counting. Since we devise a top-down parsing mechanism, we use the knowledge of supra-segmental boundaries to determine the underlying linguistic event.

To perform pattern matching, the similarity measure of binary structures must be quantified. There are many metrics formulated for this purpose (Choi and Cha, 2010) which differ mainly in the way that positive/negative match or different mismatches are addressed. We conducted thorough tests on the metrics defined in (Choi and Cha, 2010); Figure 5 compares and contrasts a few representative results.

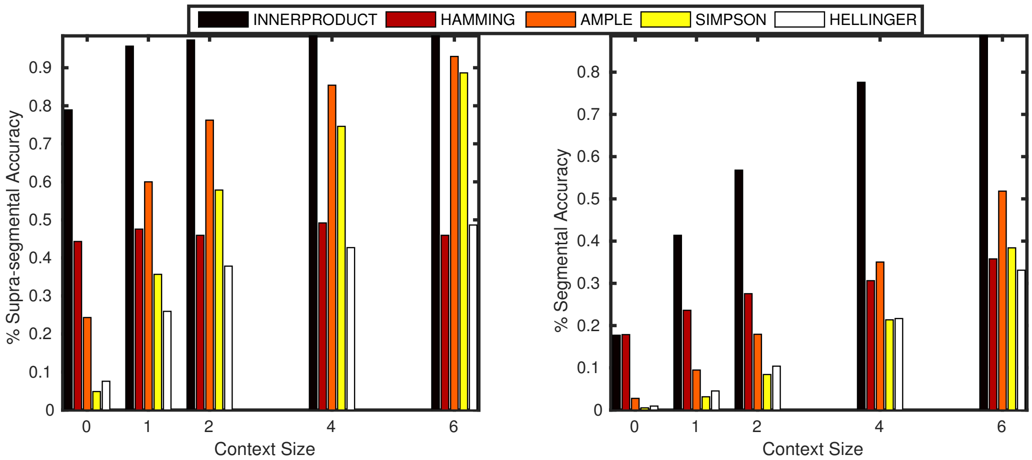


Figure 5: *Comparison of the performance of accent detection using various binary similarity measures. The measures are selected from (Choi and Cha, 2010). The results of Jaccard* (2) *is the same as innerproduct.*

We can see that the fast and simple *innerproduct* is the most effective similarity metric; it quantifies the positive and negative matches between the two binary structures. On the other hand, Hamming similarity measure that quantifies the mismatches does not perform well for linguistic parsing. The Jaccard (2) formula yields similar results to innerproduct. Hence, we choose the innerproduct for its efficiency in our linguistic parsing evaluation. Table 4 lists the accuracy of different parsing scenarios for English data provided in recordings from Nancy and French data available in SIWIS database.

Table 4: *Accuracy (%) of linguistic parsing using structured sparsity pattern matching with different context sizes. The results are evaluated on Nancy and SIWIS speech recordings. The binary similarity measure is innerproduct.*

Task / Context Size		0	1	2	4	6
C-V Detection	Nancy	53.5	83.3	88.2	93.9	96.7
	SIWIS	64.5	82.9	85.5	87.9	90.3
Stress Detection	Nancy	75.4	95.4	96.9	99.5	99.5
	SIWIS	96.9	98.5	98.5	98.5	98.5
Accent Detection	Nancy	78.4	96.8	97.3	98.4	99.5
	SIWIS	91.6	93.7	93.7	94.8	94.8

The results are averaged over 5-fold random selection of length 1000 consecutive segments. The high-order structured sparsity patterns are obtained by concatenating each segment with its adjacent segments on the right, and the context size denotes the number of extra segments concatenated. We can see that the higher order structured sparsity patterns enables more accurate linguistic parsing. It also confirms that the proposed structured sparsity principle is independent of language as well as phonological class definitions.

4.2.3. Dependency of Linguistic Events

Finally, we test the dependency between different supra-segmental attributes captured in codebook structures. Both stressed and accented syllables convey similar information on linguistic emphasis, the former denotes it at a lexical level while the latter designates it at a prosodic level. Hence, we hypothesize that the codebook constructed from stressed structures can be used for accent detection, and vice versa. Table 5 lists the accuracies using these linguistically relevant codebooks.

Table 5: *Accuracy (%) of parsing using linguistically relevant codebooks. Namely, we perform stress / accent detection using accent/ stress codebooks to study dependency of stressed and accented structures.*

Task / Context Size	0	1	2	4	6
Stress detection using accent codebooks	63.6	82.0	79.5	82.2	85.4
Accent detection using stress codebooks	65.4	80.0	79.5	82.2	85.4

We can see that a codebook constructed from either of stress/accent structures can be used for detection of the other with high accuracy. This study confirms the hypothesis that codebooks encapsulates linguistically relevant structures and demonstrates that accented structures are indeed highly correlated with the stressed structures.

5. Concluding Remarks

The theories of linguistics and cognitive neuroscience suggest that the phonological representation of speech places at the heart of speech temporal organization. We devised a methodology to quantify the phonological based supra-segmental primitives as essential building blocks for detection of various linguistic events. Our proposed approach relies on identification of structured sparsity patterns to learn class-specific codebooks characterizing different supra-segmental attributes. The experiments confirmed that indeed phonological posteriors convey supra-segmental information which is encoded in their support of active components, and these structures can be used as indicators of their higher level linguistic attributes.

In this context, we also verified that the class-specific structures of phonological posteriors is a property independent of language as well as definition of different phonological classes. In addition, it is robust to unconstrained and noisy recording conditions. Furthermore, the dependency of different linguistic properties such as stress and accent is captured in their codebooks which confirm the high correlation between their underlying structures.

This work quantified the supra-segmental events through the binary representation of posteriors. This quantification can be more accurate if multi-level discretization is considered to find a compromise between speaker and environmental variability encoded in the probabilities and the actual contribution of phonological classes.

In our future work, we plan to investigate more closely the relationship of the trajectories of the articulatory-bound phonological posterior features to the task dynamic model of inter-articulator coordination in speech (Saltzman and Munhall, 1989). This study will strengthen our knowledge about interpretation of phonological posteriors, when applied to different speech processing tasks. Applications include detection of syllable boundaries and subsequent bottom-up linguistic parsing (i.e., parsing without providing the segment boundaries as discussed by Ghitza (2011); Giraud and Poeppel (2012)), as well as phonetic posterior estimation for automatic speech recognition and synthesis systems, parametric speech coding, and automatic assessment of speech production.

6. Acknowledgment

Afsaneh Asaei is supported by funding from SNSF project on “Parsimonious Hierarchical Automatic Speech Recognition (PHASER)” grant agreement number 200021-153507.

7. References

References

Asaei, A., Cernak, M., Bourlard, H., Sep. 2015. On Compressibility of Neural Network Phonological Features for Low Bit Rate Speech Coding. In: Proc. of Interspeech. pp. 418–422.

Black, A., Taylor, P., Caley, R., 1997. The Festival Speech Synthesis System. Technical report, Human Communication Research Centre, University of Edinburgh.

Bouchard, K. E., Mesgarani, N., Johnson, K., Chang, E. F., Mar. 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495 (7441), 327–332.

Browman, C. P., Goldstein, L. M., May 1986. Towards an articulatory phonology. Phonology 3, 219–252.

Browman, C. P., Goldstein, L. M., 1988. Some Notes on Syllable Structure in Articulatory Phonology. Phonetica 45, 155–180.

Browman, C. P., Goldstein, L. M., 1989. Articulatory gestures as phonological units. Phonology 6, 201–251.

Browman, C. P., Goldstein, L. M., 1992. Articulatory phonology: An overview. Phonetica 49, 155–180.

Cernak, M., Garner, P. N., Lazaridis, A., Motlicek, P., Na, X., Jun. 2015a. Incremental Syllable-Context Phonetic Vocoding. IEEE/ACM Trans. on Audio, Speech, and Language Processing 23 (6), 1019–1030.

Cernak, M., Potard, B., Garner, P. N., Apr. 2015b. Phonological vocoding using artificial neural networks. In: Proc. of ICASSP. IEEE, pp. 4844–4848.

Choi, S.-s., Cha, S.-h., 2010. A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics, 43–48.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York, NY.

Dunn, G., Everitt, B. S., 1982. An Introduction to Mathematical Taxonomy. Cambridge University Press.

Fowler, C. A., Shankweiler, D., Studdert-Kennedy, M., Aug. 2015. Perception of the Speech Code Revisited: Speech Is Alphabetic After All. Psychological review.

Galliano, S., Geoffrois, E., Gravier, G., f. Bonastre, J., Mostefa, D., Choukri, K., 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In: In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006. pp. 315–320.

Ghitza, O., 2011. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Frontiers in psychology 2.

Giraud, A.-L. L., Poeppel, D., Apr. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nature neuroscience 15 (4), 511–517.

Goldstein, L., Fowler, C., 2003. Articulatory phonology: a phonology for public language use. In: Meyer, A., N.Schiller (Eds.), Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities. New York: Mouton, pp. 159–207.

Harris, J., Lindsey, G., 1995. The elements of phonological representation. Longman, Harlow, Essex, pp. 34–79.

Hickok, G., Poeppel, D., May 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8 (5), 393–402.

Hinton, G. E., Osindero, S., Teh, Y. W., Jul. 2006. A Fast Learning Algorithm for Deep Belief Nets. Neural Comput. 18 (7), 1527–1554.

Jakobson, R., Halle, M., 1956. Fundamentals of Language. The Hague: Mouton.

Jun, S.-A., 2005. Prosodic Typology. In: Jun, S.-A. (Ed.), Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, pp. 430–458.

Ladefoged, P., Johnson, K., Jan. 2014. A Course in Phonetics, 7th Edition. Cengage Learning.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., Schroeder, C. E., Sep. 2005. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. Journal of neurophysiology 94 (3), 1904–1911.

Lee, S., Yildirim, S., Kazemzadeh, A., Narayanan, S., 2005. An Articulatory study of emotional speech production. In: Proc. of Interspeech. pp. 497–500.

Leonard, M. K., Bouchard, K. E., Tang, C., Chang, E. F., 2015. Dynamic encoding of speech sequence probability in human temporal cortex. The Journal of Neuroscience 35 (18), 7203–7214.

Leong, V., Stone, M. A., Turner, R. E., Goswami, U., Jul. 2014. A role for amplitude modulation phase relationships in speech rhythm perception. J. Acoust. Soc. Am. 136 (1), 366–381.

Levelt, W. J. M., Aug. 1993. Speaking: From Intention to Articulation (ACL-MIT Series in Natural Language Processing). A Bradford Book.

Liberman, A. M., Whalen, D. H., May 2000. On the relation of speech to language. Trends in cognitive sciences 4 (5), 187–196.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E. F., Feb. 2014. Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science 343 (6174), 1006–1010.

Miller, G. A., Nicely, P. E., Mar. 1955. An Analysis of Perceptual Confusions Among Some English Consonants. J. Acoust. Soc. Am. 27 (2), 338–352.

Nam, H., Goldstein, L., Saltzman, E., 2009. Self-organization of syllable structure: A coupled oscillator model. In: Pellegrino, F., Marisco, E., Chitoran, I. (Eds.), Approaches to phonological complexity. Berlin, New York: Mouton de Gruyter, pp. 299–328.

Paul, D. B., Baker, J. M., 1992. The design for the wall street journal-based CSR corpus. In: Proceedings of the workshop on Speech and Natural Language. HLT '91. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 357–362.

Perennou, G., 1986. B.D.L.E.X. : A data and cognition base of spoken French. In: Proc. of ICASSP. Vol. 11. pp. 325–328.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M., Roberts, T., Nov. 2000. Auditory Cortex Accesses Phonological Categories: An MEG Mismatch Study. Journal of Cognitive Neuroscience 12 (6), 1038–1055.

Poeppel, D., Aug. 2003. The Analysis of Speech in Different Temporal Integration Windows: Cerebral Lateralization As 'Asymmetric Sampling in Time'. Speech Communication 41 (1), 245–255.

Poeppel, D., Oct. 2014. The neuroanatomic and neurophysiological infrastructure for speech and language. Current Opinion in Neurobiology 28, 142–149.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., Dec. 2011. The kaldi speech recognition toolkit. In: Proc. of ASRU. IEEE SPS, iEEE Catalog No.: CFP11SRW-USB.

Roekhaut, S., Brognaux, S., Beaufort, R., Dutoit, T., 2014. eLite-HTS: A NLP tool for French HMM-based speech synthesis. In: Proc. of Interspeech. pp. 2136–2137.

Saltzman, E. L., Munhall, K. G., 1989. A dynamical approach to gestural patterning in speech production. Ecological Psychology 1, 333–382.

Stevens, K. N., 2005. Features in Speech Perception and Lexical Access. In: Pisoni, D. B., Remez, R. E. (Eds.), The Handbook of Speech Perception. Blackwell Publishing, pp. 125–155.

Wernicke, C., 1874/1969. Bonston studies in the phylosophy of science. In: Cohen, R., Wartofsky, M. (Eds.), The symptom complex of aphasia: A psychological study on an anatomical basis. D. Reichel, Dordrecht, pp. 34–97.

Yu, D., Siniscalchi, S., Deng, L., Lee, C.-H., March 2012. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In: Proc. of ICASSP. IEEE SPS.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K., 2007. The HMM-based Speech Synthesis System Version 2.0. In: Proc. of ISCA SSW6. pp. 131–136.