Sparse Pronunciation Codes for Perceptual Phonetic Information Assessment

Afsaneh Asaei*, Milos Cernak*, Hervé Bourlard*°, Dhananjay Ram*°
*Idiap Research Institute, Martigny, Switzerland
°École Polytechnique Fédérale de Lausanne, Switzerland
Email: {afsaneh.asaei, milos.cernak, herve.bourlard, dhananjay.ram}@idiap.ch

***Abstract*—Speech is a complex signal produced by a highly constrained articulation machinery. Neuro and psycholinguistic theories assert that speech can be decomposed into molecules of structured atoms. Although characterization of the atoms is controversial, the experiments support the notion of invariant speech codes governing speech production and perception. We exploit deep neural network (DNN) invariant representation learning for probabilistic characterization of the phone attributes defined in terms of the phonological classes and known as the smallest-size perceptual categories. We cast speech perception as a channel for phoneme information transmission via the phone attributes. Structured sparse codes are identified from the phonological probabilities for natural speech pronunciation. We exploit the sparse codes in information transmission analysis for assessment of phoneme pronunciation. The linguists define a single binary phonological code per phoneme. In contrast, probabilistic estimation of the phonological classes enables us to capture large variation in structures of speech pronunciation. Hence, speech assessment may not be confined to the single expert knowledge based mapping between phoneme and phonological classes and it may be extended to multiple data-driven mappings observed in natural speech.**

I. Probabilistic Perception Channel

Phonemes are the set of unit sounds that distinguish one word from another in a particular language. The phoneme classes are denoted by an L dimensional random variable S with categorical distribution $(p_{s_1}, \ldots, p_{s_L})$ where p_{s_l} is the probability of phoneme s_l. Each phoneme possesses some phone attributes. The phone attributes are defined in terms of phonological classes that describe some properties of sound production such as [vowel], [fricative], [dental] and [labial]. For example, a sound "m" has the following attributes: [anterior], [voice] and [nasal]. The set of K phonological classes is denoted by $Q = \{q_1, \ldots, q_K\}$ where q_k is a discrete random variable taking binary values $\{0, 1\}$, with probability $p(q_k = 1) = p_{q_k}$.

The linguists define a *binary* association between phonemes and phonological classes [1]. The production theory of speech perception relies on detection of the underlying phone attributes and their unique combination to form the higher level phoneme perception. In practice however, detection of the phonological classes is not perfect. Hence, the present study relies on probabilistic characterization of the phonological classes as developed in [2].

To perform perceptual assessment, we consider speech perception as a channel having the phonological random variables at the input and phoneme random variables at the output [3]. Hence, we define z_t as the random variable which can take values of the set of phonological classes $Q = \{q_1, \ldots, q_K\}$; t indexes the temporal window. The posterior probabilities of all phonological classes $\{p(z_t = q_1|x_t), \ldots, p(z_t = q_K|x_t)\}$ are estimated by K deep neural network (DNN)s, each specifically trained to detect one of the classes from the input acoustic feature x_t [2]. Given the speech transcription, the posterior representation of the frames labelled as phoneme s_l yields $p(z_t|s_l, x_t)$. Accordingly, the posterior representation of the frames labelled as phonological class q_k yields $p(z_t|q_k, x_t)$.

The amount of information transmitted by the phone attribute q_k for perception of phoneme s_l is estimated as the multivariate mutual information expressed as follows where H denotes the entropy:

$$\begin{aligned}\mathcal{I}_k \equiv I(q_k, s_l, z_t) &= H(q_k, s_l, z_t) - H(q_k, s_l) - H(s_l, z_t) \\ &- H(q_k, z_t) + H(s_l) + H(q_k) + H(z_t)\end{aligned} \quad (1)$$

To calculate this quantity, the DNN phonological posteriors are used as follows. If the acoustic frame x_t is the result of the production of phoneme s_l, we assume that $p(x_t|z_t, s_l) = p(x_t|s_l)$; the intuition is that the physical process leading to the production of x_t is guided by s_l and the variable z_t is an abstract notion to exploit probabilistic association of the DNN outputs to all phonological classes. Hence, given the physical state of s_l, the observation x_t is independent of z_t or by Bayes theorem $p(z_t|s_l, x_t) = p(z_t|s_l)$. Similarly, $p(z_t|q_k, x_t) = p(z_t|q_k)$ are used for the joint probabilities required to calculate (1) [3].

II. Information of Pronunciation Codes

The perception of a trained speaker is sensitive to the structures underlying phone attributes during phoneme pronunciation [4]. To identify these structures, we consider binary representation of the phonological posteriors obtained via quantization [5]. The permissible structures corresponds to the indices of the non-zero components. The active components determine the posture of vocalization. Due to the constraints in articulation machinery, the binary codes are sparse. Fig. 1 illustrates an example of the structured sparsity underlying phonetic and phonological posteriors. The codes generated by vocalization are highly constrained. The linguists define unique binary codes per phoneme [1]. Probabilistic estimation of the phonological classes enables us to capture large variation in structures of speech pronunciation.

Figure 2 depicts the pronunciation assessment procedure. We identify all the unique sparsity structures from a large speech corpora and construct a codebook of permissible pronunciations [5, 6]. Speech perception operates on the principle of merging independent evidences based on the sparse pronunciation codes. We define a code associated to phoneme s_l as the set of $c_l = \{q_1, \ldots, q_{K_l}\}$ phonological classes. Following the principle of speech perception as partial recognition of independent phonological cues [7, 8], the probability of phoneme perception is calculated as independent combination of the constituting phonological class probabilities [3]. The information conveyed by the phonological code for perception of phoneme s_l is calculated as $\mathcal{I}_l = \sum_{k=1}^{K_l} \mathcal{I}_k$. The difference in the transmitted information calculated for perfect speaking and distorted pronunciation demonstrates the level and well as the major phoneme classes distorted due to imperfect pronunciation. An example result is illustrated in Fig. 3; the single expert knowledge based pronunciation code is used for this illustration.

Acknowledgement

We Acknowledge Swiss NSF funding "Parsimonious Hierarchical Automatic Speech Recognition and Query Detection" grant n. 200020-169398.

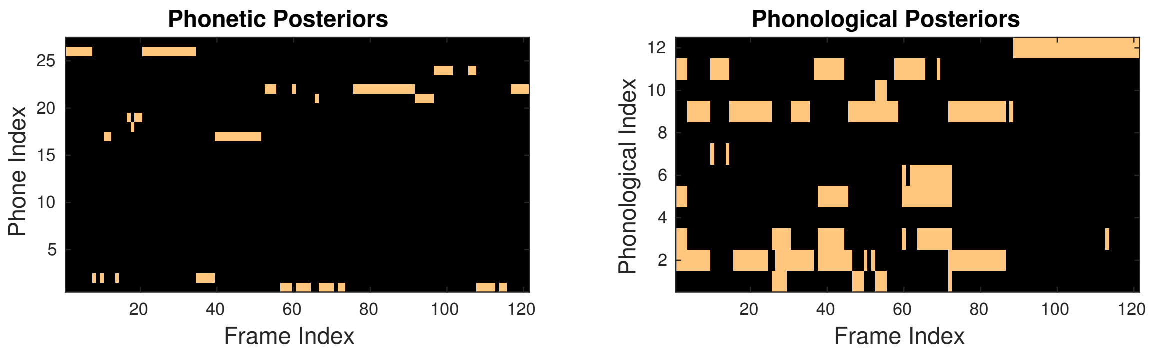

Fig. 1: Posteriograms of phonetic and phonological posteriors. We use the open-source pre-trained DNNs for estimation of posteriors [2]. Binary structures define the variants of the permissible pronunciations. The structured sparse binary codes of phonological posteriors form the codebook of permissible pronunciations. The codes generated by vocalization are highly constrained. If Q denote the number of phonological classes (e.g. $Q = 20$), 2^Q codes may be formed in theory. However, the investigation on large speech corpora of more than 100 hours of spontaneous conversational speech identifies less than 10^4 for the entire English speaking variations.

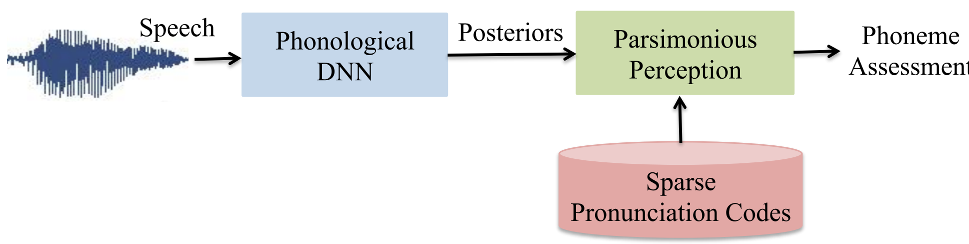

Fig. 2: Phoneme perception operates on the basis of detecting the phonological classes and merging evidences for inference of the phonemes. Linguists define a single binary phonological code per phoneme [1]. Probabilistic estimation of the phonological classes enables us to capture large variation in structures of speech pronunciation. Hence, speech assessment may not be confined to the single expert knowledge based on mapping between phoneme and phonological classes [1] and it can be extended to multiple data-driven mappings as observed in natural speech. Exploiting DNNs in probabilistic estimation of phonological classes is crucial in determining the data-driven natural pronunciation codes from phonological posteriors.

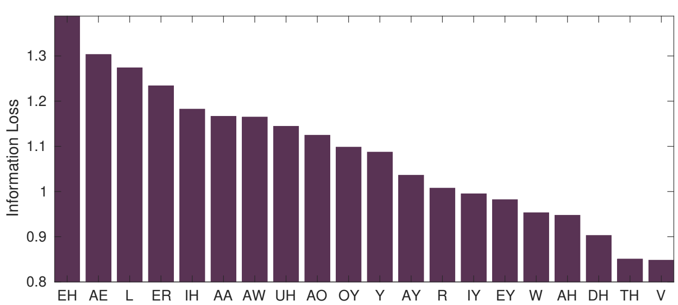

Fig. 3: Perceptual information loss due to impaired speech pronunciation demonstrated for the top 20 most affected phonemes. The phonemes are described in [9]. TORGO database of dysarthric speech is used for the experiments [10]. We can see that articulation impairment is most exhibited in pronunciation of a selective set of phonemes as recommended by the clinical tests. The source information may be analyzed at larger granularity than phonemes such as syllables. In this case, co-articulation is represented by the codebook of natural pronunciation codes.

References

[1] N. Chomsky and M. Halle, *The Sound Pattern of English*. Harper & Row, 1968.

[2] M. Cernak and P. N. Garner, "PhonVoc: A Phonetic and Phonological Vocoding Toolkit," in *Proc. of Interspeech*, 2016.

[3] A. Asaei, M. Cernak, and H. Bourlard, "Information Transmission Analysis of Production-Perception Efficiency: Case Study of Speech Pathology," Idiap-RR-30-2016. [Online]. Available: http://publications.idiap.ch/index.php/publications/show/3512

[4] C. A. Fowler, D. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code revisited: Speech is alphabetic after all," *Psychological Review*, vol. 123(2), pp. 125–150, 2015.

[5] M. Cernak, A. Asaei, and H. Bourlard, "On structured sparsity of phonological posteriors for linguistic parsing," *Speech Communication*, vol. 84, pp. 36–45, 2016.

[6] A. Asaei, M. Cernak, and H. Bourlard, "On Compressibility of Neural Network Phonological Features for Low Bit Rate Speech Coding," in *Proc. of Interspeech*, 2015, pp. 418–422.

[7] J. B. Allen, "Articulation and intelligibility," *Speech and Audio Processing Lectures, Pub: Morgan & Claypool*, ISBN-1598290088, 2005.

[8] B. Gold, N. Morgan, and D. Ellis, *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[9] R. L. Weide, "The CMU pronouncing dictionary," 1998. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[10] F. Rudzicz, A. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.