IDIAP RESEARCH REPORT

COGNITIVE SPEECH CODING

Milos Cernak Afsaneh Asaei

Idiap-RR-27-2016

Version of NOVEMBER 07, 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

Cognitive speech coding

Milos Cernak, Afsaneh Asaei

November 6, 2016

Abstract

The speech signal coveys both the linguistic-symbolic and continuous-acoustic information. The former is the result of underlying cognitive speech processes, whereas the latter is the result of motor control speech processes. The gap between the cognitive and motor speech processes is narrowing by converging of speech engineering and motor control, psycholinguistics, neuropsychology and speech neuroscience, and recent deep learning approaches.

The purpose of this paper is to propose a novel architecture of speech coding. We attribute the novel speech coding as cognitive, for compressing the speech signal into code that can be interpreted at the linguistic level, and manipulated by the computational models of speech production, such as the Directions Into Velocities of Articulators model and the Hierarchical State Feedback Control model. Linguistically relevant transmission code brings novel functionality to speech transmission systems, performing tasks such as automatic dialect correction of the speakers, or intelligibility enhancement of speakers with motor speech disorders. The proposed speech coding facilitates an integration of speech transmission with higher level sequential speech applications, such as automatic speech recognition and synthesis, and machine translation systems.

Index Terms: Speech coding, cognition, speech perception, speech production

1 Introduction

Speech coding is an essential technology in information transmission. Though the band-limited wired and wireless communication systems have changed from analogue to digital, speech coding remains the same in last 45 years, based on the waveform and Linear Predictive Coding (LPC) coding [1]. LPC coding is used in the majority of standardised higher bit-rate speech coding [2, 3], and also in lower bit-rate coding [4, 5, 6, 7, 8, 9].

Historically, the main purpose of speech coding is speech compression [3] for transmission of the speech signal by limited-band telecommunication channels. The speech signal coveys both the linguistic-symbolic and continuous-acoustic information. The former is the result of underlying cognitive speech processes, whereas the latter is the result of motor speech processes. The gap between the cognitive and motor speech processes is narrowing by recent deep learning approaches and converging of speech engineering and neuroscience.

The purpose of this paper is to propose a novel architecture of speech coding where the speech coders “understand” the transmitted speech code that is linguistically relevant. Linguistically relevant transmission code could bring novel functionality to speech transmission systems, performing tasks such as automatic:

1. Dialect correction.
2. Pronunciation improvements for people with phonological and articulatory disorders.
3. Intelligibility enhancements.

The first two tasks are performed in the transmitter, and the third task is performed in the receiver using the environment analysis.

Our motivation for devising of this emerging field comes partially from the radio communication. *Cognitive radio* emerged in the last decade [10, 11] as the means of injecting human intelligence into the telecommunication networks. Main feature is analysis of radio environment, and re-configuration of the transmission network to improve its quality and efficiency. Real-time reconfigurability is provided by software-defined radio. Similarly, we attribute the novel speech coding as *cognitive*, for compressing the speech signal into the code that can be interpreted at the linguistic level. Relation of cognitive radio and cognitive speech coding can be simply described as cognitive radio being an intelligent reconfigurable high-way, whereas cognitive speech coding are intelligent vehicles on the roads. Moreover, while cognition in cognitive radio is taken as artificial computational mechanism applied to radio transmission, cognition in cognitive speech coding is much more integrated within the speech transmission itself by using computational models of human speech production and perception. Cognition in Figure 1 performs feedback to conventional feedforward speech coding, and corrects the transmitted speech code.

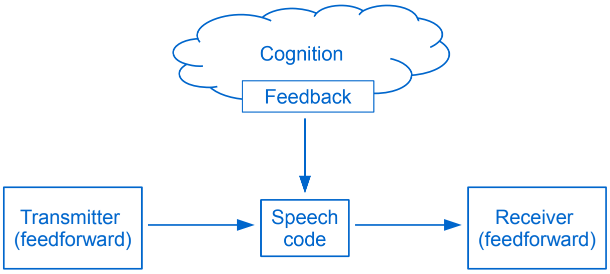

Figure 1: The design of cognitive speech coding. Feedforward speech coding is complemented by cognitive feedback.

LPC coding already uses models of human auditory perception, such as perceptual weighting of the residual quantization error and adaptive postfiltering [12], in order to minimise different types of auditory distortion. And the LPC model itself is the speech production model. Also perceptual audio coders, well represented by the MPEG (moving picture experts group) standards, use perceptual limitations of the human ear to encode arbitrary signals. A key difference with our proposal is in using cognitive targets that extend the auditory targets specified by human auditory perception properties with linguistically relevant speech targets.

The devised technology bridges speech engineering, acoustic phonetics, phonology and speech neuroscience. This paper outlines the cognitive aspects of the proposed speech coding scheme, put it in the context of mentioned research fields as an interdisciplinary study, and describes new functionality that it offers.

2 An outline of the proposed paper

2.1 Speech code

Speech communication can be defined as looped speech perception and production. Most of speech perception and production theories define an invariant speech representation that exists between the cognitive (linguistic) processes and the motor control (articulatory) processes. Let us call the invariant speech representation as speech code that lies in the intersection of the cognitive and motor control processes. Speech code is greatly debated in motor control, psycholinguistics, neuropsychology and speech neuroscience, and recent findings suggest that speech code includes vocal tract shapes, time-varying articulatory gestures [13, 14, 15, 16], and auditory and somatosensory targets [17]. Speech code can be defined at the phonetic or phonological level, and its basic units are usually syllables.

Cognitive processes consist of feedforward and feedback processes. For example, the Directions Into Velocities of Articulators (DIVA) model contains a feedback control map [18], where speech acquisition and production processes are influenced by auditory feedback (sound perception) and somatosensory feedback (tactile feedback, for example, if the tip of the tongue has touched the alveolar ridge during the [t] sound production). In addition, the Hierarchical State Feedback Control (HSFC) model[19], which posits internal error detection and correction processes, can detect and correct speech production errors prior to articulation.

Figure 2 considers the speech code as the invariant speech representation used in *cognitive speech coding*. Tough the figure shows only human speech production process, speech production and perception share the same set of invariants, in both Motor Theory [20] and Direct Realist Theory [21] of speech perception. Cognitive speech coding is thus defined as:

> “Cognitive speech coding is the speech transmission or storage system based on invariant articulatory gestural, auditory and somatosensory targets, which are used in the cognitive

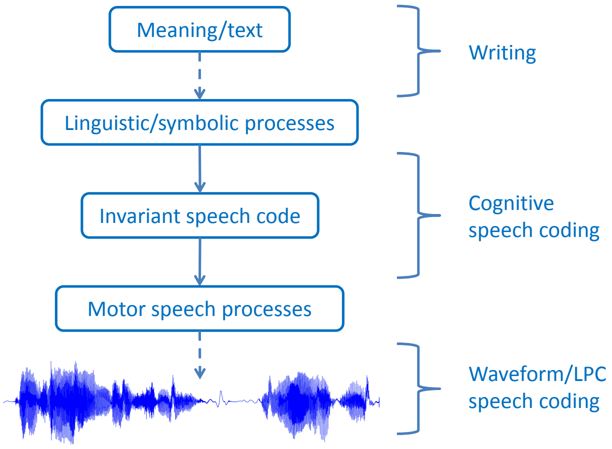

Figure 2: Simplified human speech production process. Writing is a transmission of discrete linguistic information, whereas LPC coding is a transmission of the continuous speech signal. Speech code as the invariant speech representation that lies between cognition and articulation, and speech perception and production, is an information unit transmitted in cognitive speech coding.

> tasks as the speech representation in the speech perception and production models, built within the coder."

2.2 Cognitive speech coding

We assume that cognitive speech coding is based on neural networks (NNs), because the DIVA and HSFC computational models are neural network based, and also deep learning in general has influenced recent speech coding development. For example, NN based speech coder can be realized as a composition of deep and spiking NNs [22]. The deep neutral networks encode and decode the speech signal based on the phonological speech representation that is partially related to articulatory gestures[23], and the spiking NN is used for prosody encoding.

Figure 3 shows cognitive speech coding built as an extension of NN speech coding with formalized feedback processing performed by the cognitive tasks:

1. *Analysis feedback*: the analysis feedback is performed in the transmitter.
2. *Synthesis feedback*: the synthesis feedback is performed in the receiver.

Both cognitive tasks share the Universal Perception/Production Targets (UPPTs).

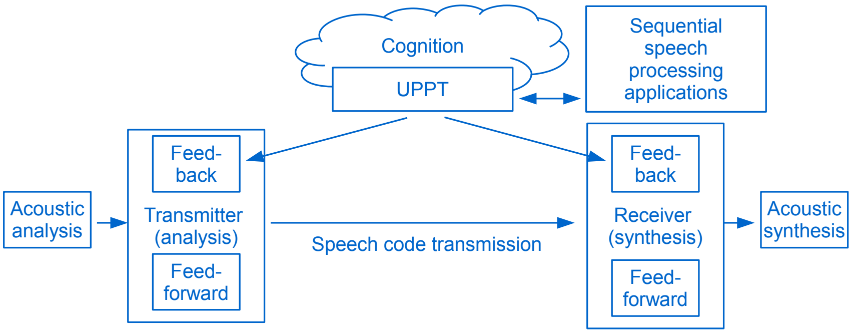

Figure 3: Building blocks of cognitive speech coding. UPPT stands for Universal Perception/Production Targets. The UPPTs can be defined at the phonetic or syllabic level, functionally identical to the speech map defined by the DIVA model[18]. Eventually, the UPPTs can be accessed via an API by any sequential speech processing application that could be in future integrated with cognitive speech coding.

2.3 Cognitive tasks

Cognitive tasks model the feedback in speech perception and production. The feedback in general consists of articulatory gestural feedback and auditory feedback. The latter is broadly related to auditory targets used in non-intrusive estimation of speech quality, such as the ITU-T recommendation P.563 that defines a perceptual model of speech, and it is not discussed here. Rather, we describe novel feedback for the speech code constructed from the articulatory gestures, currently not used by waveform/LPC coding.

2.3.1 Analysis feedback

The feedforward processing transforms the acoustic feature observation sequence $X = \{\vec{x}_1, \ldots, \vec{x}_n, \ldots, \vec{x}_N\}$, where N denotes the number of segments in the utterance, into the speech code, a parameter sequence $\vec{z}_n = [z_n^1, \ldots, z_n^K, \ldots, z_n^k]^\top$, where the n-th frame consists of posterior probabilities $z_n^k = p(c_p|x_n)$ of K classes, and $.^\top$ is the transpose operator.

The analysis feedback corrects the speaker-dependent deviations of the parameter sequence Z using the UPPTs. The UPPTs is a parameter sequence $U = \{\vec{u}_1, \ldots, \vec{u}_m, \ldots, \vec{u}_M\}$, where M denotes the length of the sequence. Using of the U sequence is conceptually similar to using of the excitation code-book in analysis-by-synthesis LPC coding. Figure 4 shows the cognitive speech coder where the feedback correction is based on error minimization of the feedforward sequence Z and the target sequence U. Error minimization can perform error correction as defined by the DIVA and HSFC models.

Let us consider the task of dialect correction. Error minimization might be implemented as an automatic accentedness evaluation of non-native speech [24, 25] in an closed-loop fashion.

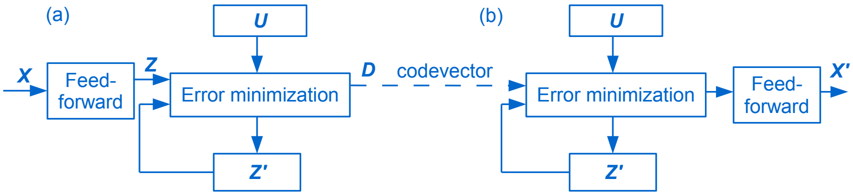

Figure 4: The feedforward and feedback processing of the cognitive speech encoder (a) and decoder (b). U denotes the UPPT parameter sequence, Z' is the candidate corrected sequence, and D is the quantized sequence.

2.3.2 Synthesis feedback

Receiving sequence D is the codeword of the quantized Z sequence, distorted by environmental noise. The role of error minimization in the decoder is thus recovering of the original speech code Z. The synthesis feedback can perform intelligibility enhancement; for example, by a transformation introduced in [26]. The transformation could be realized by articulatory gestural and auditory feedback using the UPPTs.

Once Z sequence is recovered, it is re-synthesized by acoustic synthesis. More than 77% of all speech degradation in NN speech coding comes from the parametric vocoding [22]. Therefore, re-synthesis of raw speech samples using a fully convolutional NN[1] is a promising synthesis technique to be used in cognitive speech coding.

2.4 Security issues

Similarly as in cognitive radio and software-defined networking, security issues have to be considered. Cracking the speech signal into the invariant speech code results into breaking up of the speaker-dependent and speaker-independent information transmission. The anonymous "content" of the communication might thus facilitates spoofing attacks of biometric systems. Also, the speaker identity is transmitted, and thus such transmission has to be well encrypted.

Hence, we propose two versions of cognitive speech coding: (i) "hard" cognitive speech coding, and (ii) "soft" cognitive speech coding. The former coding is the cognitive speech coding that transmits the

[1] WaveNet: A Generative Model for Raw Audio, https://arxiv.org/pdf/1609.03499.pdf

speech code incl. the speaker identity directly. The latter coding is a hybrid system of current waveform/LPC and cognitive coding, where only the analysis-by-synthesis blocks of waveform/LPC coding are replaced by the cognitive analysis and synthesis feedbacks, and the transmitted information is untouched.

References

[1] B. S. Atal, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, pp. 154–161, Mar. 2006.

[2] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache, Y. Kamamoto, K. Kikuiri, S. Ragot, J. Faure, H. Ehara, V. Rajendran, V. Atti, H. Sung, E. Oh, H. Yuan, and C. Zhu, "Overview of the EVS codec architecture," in *Proc. of ICASSP*, pp. 5698–5702, IEEE, Apr. 2015.

[3] J. Gibson, "Speech Compression," *Information*, vol. 7, pp. 32+, June 2016.

[4] D. Wong, B.-H. Juang, and A. Gray, "An 800 bit/s vector quantization LPC vocoder," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 770–780, Oct. 1982.

[5] S. Roucos, R. Schwartz, and J. Makhoul, "Segment quantization for very-low-rate speech coding," in *Proc. of ICASSP*, vol. 7, pp. 1565–1568, IEEE, May 1982.

[6] S. Roucos, R. Schwartz, and J. Makhoul, "A segment vocoder at 150 b/s," in *Proc. of ICASSP*, vol. 8, pp. 61–64, IEEE, Apr. 1983.

[7] D. Wong, B. Juang, and D. Cheng, "Very low data rate speech compression with LPC vector and matrix quantization," in *Proc. of ICASSP*, vol. 8, pp. 65–68, IEEE, Apr. 1983.

[8] C. Tsao and R. Gray, "Matrix quantizer design for LPC speech using the generalized Llyod algorithm," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 33, pp. 537–545, June 1985.

[9] Y. Shiraki and M. Honda, "LPC speech coding based on variable-length segment quantization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 36, pp. 1437–1444, Sept. 1988.

[10] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, pp. 13–18, Aug. 1999.

[11] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, pp. 201–220, Feb. 2005.

[12] M. Hasegawa-Johnson and A. Alwan, *Wiley Encyclopedia of Telecommunications*, ch. Speech coding: fundamentals and application. John Wiley & Sons, Inc., 2003.

[13] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonology*, vol. 3, pp. 219–252, May 1986.

[14] C. P. Browman and L. M. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, pp. 201–251, 1989.

[15] C. P. Browman and L. M. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[16] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation.," *Nature*, vol. 495, pp. 327–332, Mar. 2013.

[17] F. H. Guenther and G. Hickok, *Role of the auditory system in speech production*, vol. 129, pp. 161–175. Elsevier, 2015.

[18] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and Cognitive Processes*, vol. 26, pp. 952–981, Aug. 2011.

[19] G. Hickok, "The architecture of speech production and the role of the phoneme in speech processing," *Language, Cognition and Neuroscience*, vol. 29, pp. 2–20, Jan. 2014.

[20] A. M. Liberman and I. G. Mattingly, “The motor theory of speech perception revised,” *Cognition*, pp. 1–36, 1985.

[21] C. A. Fowler, D. Shankweiler, and M. Studdert-Kennedy, “Perception of the Speech Code Revisited: Speech Is Alphabetic After All.,” *Psychological review*, Aug. 2015.

[22] M. Cernak, A. Lazaridis, A. Asaei, and P. N. Garner, “Composition of Deep and Spiking Neural Networks for Very Low Bit Rate Speech Coding,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 2301–2312, Dec 2016.

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

[24] R. Rasipuram, M. Cernak, A. Nachen, and M. Magimai-Doss, “Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities,” in *Proc. of Interspeech*, pp. 648–652, 2015.

[25] R. Rasipuram, M. Cernak, and M. Magimai-Doss, “HMM-based Non-native Accent Assessment using Posterior Features ,” in *Proc. of Interspeech*, pp. 3137–3141, 2016.

[26] F. Rudzicz, “Adjusting dysarthric speech signals to be more intelligible,” *Computer Speech & Language*, vol. 27, pp. 1163–1177, Sept. 2013.