Cognitive Speech Coding

Milos Cernak, *Senior Member, IEEE,* Afsaneh Asaei, *Senior Member, IEEE,* Alexandre Hyafil

***Abstract*—Speech coding is a field where compression paradigms have not changed in the last 30 years. The speech signals are most commonly encoded with compression methods that have roots in Linear Predictive theory dating back to the early 1940s. This paper tries to bridge this influential theory with recent cognitive studies applicable in speech communication engineering.**

This tutorial article reviews the mechanisms of speech perception that lead to perceptual speech coding. Then it focuses on human speech communication and machine learning, and application of cognitive speech processing in speech compression that presents a paradigm shift from perceptual (auditory) speech processing towards cognitive (auditory plus cortical) speech processing. The objective of this tutorial is to provide an overview of the impact of cognitive speech processing on speech compression and discuss challenges faced in this interdisciplinary speech processing field. In this context, it covers the traditional speech coding techniques as well as emerging approaches facilitated by deep learning computational methods. The tutorial points out key references on fundamental teachings of psycholinguistics and speech neuroscience and provides a valuable background to beginners and practitioners on the promising directions of incorporating principles of cognitive speech processing in speech compression.

***Index Terms*—Speech coding, speech production and perception, cognition, deep learning**

I. Introduction

Speech coding is an essential technology in information transmission and communication systems. Human cognitive processing operates at about $50\,\mathrm{bps}$ (bits per second), which corresponds roughly to the speech production semantics as the rate of phonemic information in speech (e.g., most languages have approximately 32 phonemes, encoded with 5 bits, and 1 s of speech has perhaps 10 phonemes), and the sensory system is known to encode non-redundant structures [1]. Efficient coding maximizes the amount of information conveyed about the sensory signal to the rest of the brain. The incoming acoustic signal is transmitted mechanically to the inner ear and undergoes a highly complex transformation before it is encoded efficiently by spikes at the auditory nerve. This great efficiency in information representation has inspired speech engineers to incorporate aspects of cognitive processing in when developing efficient speech technologies.

Speech coding is a field where research has slowed considerably in recent years. This has occurred not because it has achieved the ultimate in minimizing bit rate for transparent speech quality, but because recent improvements have been small and commercial applications (e.g., cell phones) have been mostly satisfactory for the general public, and the growth of available bandwidth has reduced requirements to compress speech even further. However, better compression is always welcomed, e.g. in large archival systems, etc. This article presents an overview of the basics of speech representation and speech neuroscience, and outlines cognitively inspired speech coding that promises higher compression, adaptability and robustness of the next generation of speech coding technology.

Historically, the mechanisms of speech perception lead to perceptual speech coding [2], primarily suitable for digital audio. Substantial progress in this context incorporates mechanisms to "optimize" coder performance for the human ear in the context of sub-band (transform) coders. On the other hand, the most common speech coding for medium to low bit rates is based on models of human speech production, realized as linear predictive vocoders, and analysis-by-synthesis linear predictive coders. Unified audio and speech coding is usually realized with real-time switching according to the input signal type. For example, the Enhanced Voice Services (EVS) coder standardized in 2015 by 3GPP offers new features and improvements for low-delay real-time communication systems, higher quality for clean/noisy speech, mixed content and music, including support for wideband, super-wideband and full-band content [3]. However, the core speech compression method is Algebraic Code Excited Linear Prediction (ACELP) proposed in 1987 by Adoul et al. [4]. The compression paradigm thus did not change significantly in the last 30 years, and has its roots in Linear Predictive theory dating back to the early 1940s [5], [6]. The significance of this tutorial is in bridging this influential

Milos Cernak and Afsaneh Asaei are with Idiap Research Institute, Centre du Parc, Rue Marconi 19, 1920 Martigny, Switzerland. Email: milos.cernak@ieee.org, afsaneh.asaei@idiap.ch.

Alexandre Hyafil is with the Center for Brain and Cognition at Universitat Pompeu Babra, Barcelona, Spain. Email: alexandre.hyafil@gmail.com

theory with cognitive studies highlight applicability to speech communication engineering.

Research studies during the last decade incorporate additional aspects of speech processing, namely functional and temporal organization of human speech and language processing. Figure 1 shows the overall speech perception process. Speech signal is treated indiscriminately by subcortical structures from other types of acoustic input. Such processes have now been fairly well characterized, particularly the cochlea where it is decomposed into different frequency channels forming, what has been coined, the auditory spectrogram [7]. Automatic speech processing is much less inspired from the subsequent stages in the auditory cortex, and in particularly omits how this continuous representation is transformed into a discrete representation, i.e., a lexicon. Such a transformation is particularly difficult because the different units (phonemes, syllables, etc.) have varying durations and operate at different time scales.

The cognitive speech processing introduced in this article, is specifically based on a dual-stream cortical circuit [8], and it presents a paradigm shift from perceptual (auditory) speech processing towards cognitive (auditory/peripheral+cortical/central) sparse speech processing. Perception is extensively studied in general auditory and speech processing [9]. Biologically, not only the cochlea but also the auditory cortex contribute to speech perception. Auditory sensations reach perception only if received and processed by a cortical area.

Cognitive speech processing covers in particular the temporal aspects of speech processing. As shown by Fig. 1, the auditory cortex must deal with the different time scales pertaining to speech, and so one prominent hypothesis is that speech is first parsed into chunks corresponding to syllables and phonemes and then each chunk is categorized [10]. It has been shown that, during speech processing, the brain generates a cortical oscillation in the θ-range (3-8 Hz) that may correspond to the syllable rate, and faster γ-range oscillations (25-40 Hz) that correspond to the more transient acoustic properties. As a result, the fine (phonetic) structure of the speech (the energy bursts underlying consonants) have signal modulation even above $40\,\mathrm{Hz}$. Although psychology and speech engineering already have a long history of competing theories of speech perception [11], recent experimental and theoretical developments in neuroscience support the idea that this cortical temporal sampling is thought to play a key role in human speech processing [12].

The principle of this multi-resolution temporal sampling has been studied in the context of speech compression. The basic idea is based on packaging information into units of different temporal granularity, such as phonemes and syllables, in parallel. As an example, the incremental phonetic vocoder – cascaded speech recognition and synthesis systems – extended with syllable-based asynchronous information transmission mechanisms was recently proposed [13]. The principles of asynchronous processing are fundamental in cortical perception processing; asynchronicity exists in visual perception [14], in audiovisual perception [15], and in asynchronous evolution of various articulatory feature streams of speech recognition [16].

The objective of this tutorial is to provide an overview of the impact of cognitive speech processing on speech compression, and outline challenges faced in this interdisciplinary signal processing field. The article relates to recent findings of speech and language neuroscience with traditional speech coding techniques in the context of recent deep learning computational methods. The tutorial points out key references on fundamental teachings of psycholinguistics and neurolinguistics and provides a valuable background on the promising directions to incorporate principles of cognitive speech processing in speech compression.

Cognition in this article is assumed to be information processing in the central nervous system, after the peripheral auditory system. We avoid any broader definition that relates to abstract concepts such as memory, meaning, mind and intelligence. Although cognition in speech engineering might be understood as speech perception, for clarity of our presentation, Section II assumes cognition as the underlying processes (algorithms) existing in both speech perception and production. Section III briefly reviews perceptual audio coding and linear prediction speech coding in order to provide association with cognitive speech processing. The last Section IV introduces cognitive speech coding and compares its properties with linear predictive coding, concluding by outlining the challenges faced in this interdisciplinary speech compression field.

II. Human Cognitive Speech Processing

This section reviews the key results of encoding of sound, and in particular speech sounds, by humans, and its sparse and cortical representations. It is focused on the findings that have impact on speech processing systems.

A. Human speech coding

Historically, we learned about the functional anatomy of speech and language by observing its dysfunction. In the late 19th century, Wernicke observed that fluent aphasia (patients who utter fluent but meaningless

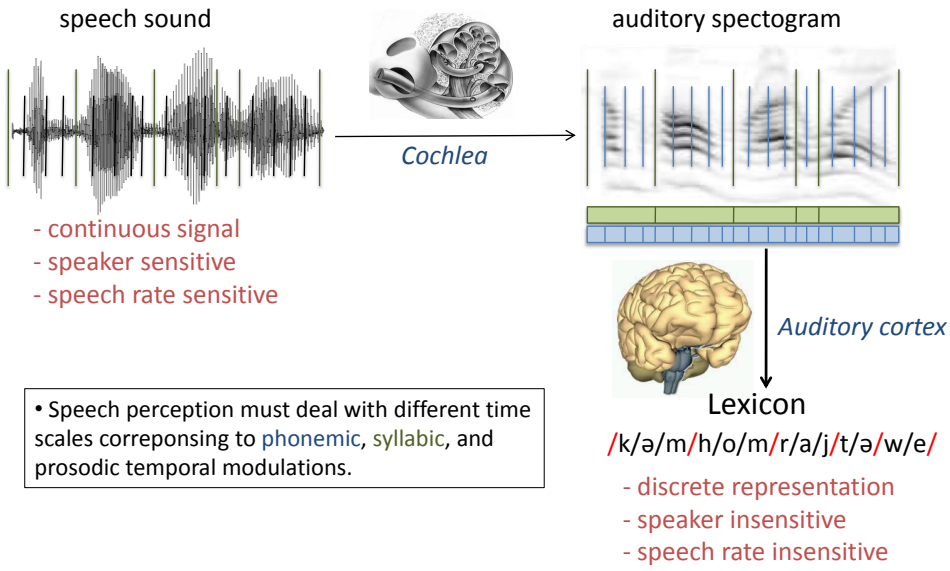

Fig. 1. Overall speech perception process carried out by the peripheral (cochlea) and the central auditory (primary auditory cortex) systems. The green and blue boxes with corresponding vertically-spaced lines represent syllabic and phonetic speech segmentation, respectively.

speech, with impaired comprehension) was associated with damage in the superior temporal gyrus (STG) [8], [17]. On the other hand, damage to Broca's area causes non-fluent aphasia, which results in intact comprehension but partial loss of the ability to produce written and spoken language [18]. The first stages of acoustic treatment (the peripheral auditory system, the front-end), before reaching the auditory cortex, have been well characterized. This process is summarized in a popular computational model [7], which lists pre-cortical steps as the following: i) a decomposition of the acoustic signal into filter banks through wavelet transform in the cochlea; ii) a high-pass filter, non-linear compression and low-pass filtering by hair cells in the auditory nerve; iii) an enhancement of the frequency selectivity (and rectification) through a lateral inhibitory network in the cochlear nucleus; an finally, iv) a further low-pass filtering in the midbrain. Subsequent stages in auditory perception involving auditory cortex (spreading from primary auditory cortex, A1 , i.e. the part of auditory cortex that receives direct input from the thalamus) and other cortices are the subject of current intense research. These are the main brain areas where cognitive speech processing takes place.

Human auditory coding evolved into highly efficient coding strategies to maximize the information conveyed to the brain (and between brain areas) while minimizing the required energy and neural resources [19]. Sparse coding scheme and hierarchical processing are central to A1 information extraction and transformation, and are present from peripheral to central auditory structures [12], [20], [21]. In turn, the principles of sparse and hierarchical (deep) structures in representation learning of sound has led to advancements in speech processing techniques [22].

Investigations into electrophysiological recordings show that no more than 5% of neurons of A1 fire above 20 spikes/s in response to acoustic stimulation. This observation suggests that the auditory responses are "sparse" and highly selective [23], which permits more accurate representations and a better discrimination of auditory stimuli [24].

1) A Dual-Stream Model: Not only speech (production and perception) data but also promising functional neural data from the brain activity during speech are increasingly used to devise cognitive models of speech and language production and perception. Figure 2 shows one prominent example, the simplified dual-stream cortical circuit linking cortical network architecture with speech processing, that leads to a different paradigm in cognitive speech coding. The first cortical stages of auditory processing take place in the auditory cortex and more anterior parts of the STG. Spectrotemporal analysis on the pre-cortical input allows unveiling spectrotemporal patterns (formants, place of articulation, etc.) and decode the associated phonemes [27], and further phonological-level processing [25]. Damage of this brain area results for example in speech agnosia, known as an incapability to comprehend spoken words despite intact hearing, speech production, and reading ability.

The dual-stream model posits two diverging pathways emerging from the auditory cortex: along the ventral

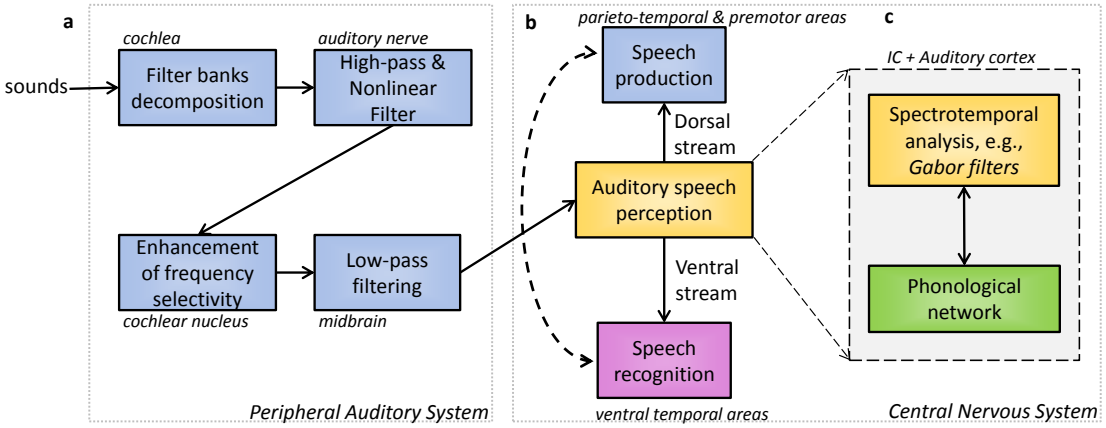

Fig. 2. Simplified functional model of human speech processing. **a)** The first stages of acoustic treatment – peripheral auditory system. **b)** The dual-stream model of human speech and language processing [8]. Auditory sensations are processed by primary auditory cortex in two processing pathways 1) the dorsal (back or posterior) stream for speech production and 2) the ventral (front or anterior) stream for speech recognition and understanding. **c)** Functional decomposition of primary auditory cortex [25]. The processing consists of spectrotemporal and phonological encoding. We can formalize phonological coding for example by articulatory phonology [26] that aims to unify low- and high-dimensional description of a single (speech) system.

stream, including ventral regions of the temporal cortex, sounds are mapped to meaning using underlying phonological representations; along the dorsal stream, including parieto-temporal as well as premotor areas, speech sounds are associated with articulatory patterns, a direct link between perceptual and motor representations of speech based again on the phonological speech representation. The former stream supports speech recognition and understanding, whereas the latter stream reflects beneficial effects of auditory feedback on speech production. While the discrete phonological speech representation plays a fundamental role in both speech perception and production [28], at the level of the temporal lobe, current understanding favors a hybrid continuous/discrete nature of speech cognition that accommodates the relevance of the continuous aspects of speech for explaining certain features of speech (e.g., prosody).

2) Temporal Organization of Speech Perception: Speech remains mostly intelligible when the spectral content is replaced by noise, and only the envelope – modulation of signal energy – is preserved, especially all modulations below 12 Hz [29], [30]. Temporal regularities of the speech signal thus play a key role in speech processing. Based on psychoacoustics and neuroimaging studies, researchers proposed that intrinsic neural oscillations play a special role in segmenting speech along time scales of distinct granularities [10]. Evidence suggests that the auditory cortex segregates acoustic information 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) [31], with a strong asymmetry between left and right hemispheres [32].

Functional organization of ventral sensorimotor cortex supports the 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 [33]. Building upon temporal information segregation in auditory cortex, Figure 3 draws a structural and temporal organization of a bottom-up organization of human speech perception.

Fig. 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 [32].

3) Top-down control and speech predictions: There is a long-standing debate about the functional significance of the massive feedback projections from higher areas to lower areas in the cortex. Indeed, most cognitive processes rely on both feedforward (bottom-up) and feedback (top-down) operations. One possible role for these projections would be to convey predictions about the upcoming sensory information. A neural signature of such syllabic predictions was uncovered using Magneto-EncephaloGraphy (MEG) [34]. On the computational side, influential models of speech perception dating back to the 1970s have proposed strongly opposing views on whether lexical predictions can bias pre-lexical

perceptual decisions. More recently, these models have been recast within a Bayesian framework, whereby the brain computes posterior probabilities about phonemes, syllables and/or lexical units (see [35], [36] for recent overviews). According to the predictive coding theory, bottom-up signals would only transmit the error between the predictions and the actual sensory information [37]. This would allow massive reduction in the size of the information passed on from sensory to higher areas, especially in the context of speech, which contains many redundancies at various levels (see next section about sparse coding).

Speech acquisition and production models such as the Directions Into Velocities of Articulators (DIVA) [38] contain auditory feedback and somatosensory (tactile) feedback, for example, if the tip of the tongue has touched correctly the alveolar ridge during the [t] sound production. Another speech production model, the Hierarchical State Feedback Control (HSFC) model that posits internal error detection and correction processes, can in addition detect and correct speech production errors prior to articulation (see [39] for detailed review).

B. Biologically inspired speech representations

Biologically inspired systems have been proposed to enhance the spectral representations of speech. The proposed models approximate the peripheral and the central auditory systems, with high-dimensional short-time vectors (for example, vectors 3840 long in [40], and 6766 in [41]). In this section we introduce recent cortical representations of speech, used for automatic speech syllabification [42], speech recognition [43] and voice activity detection [44].

1) Perception of noise intrusiveness: For natural audio signals like speech and environmental sounds, gammatone atoms have been derived as expansion functions that generate a nearly optimal sparse signal model [45]. Furthermore, gammatone functions are established models for the human auditory filters employed in the cochlea. Recent advances exploit this property in developing a sparse gammatone signal model that can predict the annoyance of background noise in listening to the speech signals as perceived by humans. This study demonstrates that the number of gammatones required to encode the noise is directly correlated with the perception of noise intrusiveness [46].

2) Spectrotemporal features: Sparse representation is found useful in learning structures in the spectrogram representation of sound such as harmonics, formants, onsets and localized patterns [47]. These sparse acoustic features resemble neuronal receptive fields reported in the Inferior Colliculus (IC), as well as auditory thalamus and cortex, and sparse modeling of neurons exhibits the same tradeoff in spectrotemporal resolution as has been observed in IC. This model is able to predict the receptive fields of neurons in the ascending mammalian auditory pathway beyond the auditory nerve [47]. Finally, within the central auditory system, tuning properties of auditory neurons in A1 are well described by sparse spectro-temporal filters, again consistent with sparse encoding of acoustic information [48].

Psychoacoustical and neurophysiological results indicate that spectrotemporal modulations play an important role in sound perception. Speech signals, in particular, exhibit distinct spectrotemporal patterns that are well matched by receptive fields of cortical neurons. Hence, methods that can capture spectro-temporal modulations are considered to improve the performance of the speech recognition systems. Along this line the Gabor shaped localized spectrotemporal features were extensively deployed by scientists and engineers at Berkeley for robust speech recognition systems[1]. The Gabor filters can model the shape of receptive fields of cortical neurons in the primary auditory cortex [49], [50]. The localized features are obtained by 2D convolution of an auditory (mel) spectrogram with the Gabor filters.

The Gabor filters are defined as the product of a complex sinusoidal function $s(n, k)$ with n and k denoting the time and frequency index, and a short-time window function $w(n, k)$. Spectrotemporal Gabor features may improve recognition results in all acoustic conditions. For example, automatic speech recognition in one-speaker conditions with reverberation and noise resulted in large relative Word Error Rate improvements of at least 52% [43].

3) Temporal encoding systems: A computational model [51] of self-generated neural oscillations showed as a proof-of-concept that: (i) such neural oscillations can reliably signal syllable boundaries and that (ii) detected syllable boundaries can improve recognition of linguistic units in a parallel neural pathway. In such a model, coupled excitatory and inhibitory neurons intrinsically synchronize around $6\,\mathrm{Hz}$, and automatically lock to edges in speech amplitude that convey the syllabic flow.

The model is based on an interconnected network of leaky integrate-and-fire neurons, in essence, the network of 10 excitatory and 10 inhibitory neurons. Neural oscillations automatically lock to speech slow fluctuations that convey the syllabic rhythm: a putative syllable boundary is declared for each inhibitory spike burst, that is whenever there were at least 2 inhibitory spikes oc-

[1] http://www1.icsi.berkeley.edu/Speech/papers/gabor/

curring within a window of 15 ms grossly corresponding to the time scale of integration of cortical neurons.

C. Phonological features

Linguistic and neurocognitive studies recognize the phonological features as the essential and invariant representations used in speech temporal organization. Cernak et al. [52] hypothesized that phonological speech representation inferred using a deep learning approach can form the basis of information flow in the phonological network of the dual-stream model showed in Fig. 2. The phonological posterior probabilities estimated by a feed-forward neural network convey information at multiple temporal scales directly mapping to the syllabic and stress information.

Phonological features, known also as distinctive and phone-attribute features, are considered as a lower-dimensional – structural – representation of phonetic features, analogically to an RGB color model in which red, green and blue light is added together to reproduce colors. Articulatory phonology aims to unify the low- (abstract) and high-dimensional (physical) description of a speech system, in which lower dimensional articulatory gestures are linguistically relevant. Bouchard et al. [33] also claim that functional organisation of the ventral sensorimotor cortex supports the gestural model developed in articulatory phonology.

In the context of speech coding, the short-term speech representation inferred from the speech signal using a deep learning approach, a vector of phonological posterior features, is shown to enable high compressibility [53] and considered to be partially related to articulatory gestures, and thus to phonological processing performed in the STG. The hypothesis of the correspondence of the phonological posteriors to the gestural trajectories is also motivated by the analogy to the constriction dynamics model [54] that takes gestural scores as input and generates articulator trajectories and acoustic output. Alternatively to this constriction dynamics model, acoustic output can be generated using a phonological synthesis described in [55].

III. Audio and Speech Coding

In this section, we briefly review perceptual audio coding and linear prediction speech coding in order to provide association with the previous section. Readers can find excellent complete reviews, for example, in [56]–[59].

Figure 4 shows machine speech coding as the analogy to the dual-stream human speech coding model. The transmitted code consists of short-term filterbank

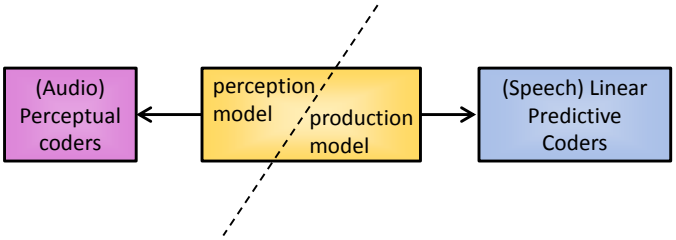

Fig. 4. Underlying models of audio and speech machine coding. Similarly to human speech coding shown at Figure 2, also two disctinct pathways (technologies) exist: perceptual-model (audio) and production-model based (speech) coders.

parameters (in perceptual coding) or linear predictive parameters (in speech coding), and long-term (prosodic) parameters. Temporal parameters are not encoded directly and phonological analysis is not performed at all. We can also observe two processing (technological) pathways, split according to the underlying human perceptual or production model. There is only a small overlap in both technological pathways leading to two distinct classes of speech coding: perceptual for audio sources and linear predictive coders for speech sources. There is currently no quest for a joint or universal coding scheme, instead both coding classes are sophistically switching in real-time according to the source type. This hybrid approach currently offers the best audio and speech coding available [60].

A. Perception-Model Based Compression

The early attempts to incorporate human speech perception into speech compression were in modeling of auditory masking performed by the inner ear and the cochlea. Modeling of dynamic masking, associated with cochlear outer-hair-cell processing [61], resulted in development of μ-law and A-law speech compression algorithms, standardized by the ITU-T G.711 standard released in 1972. These dynamic range speech compression algorithms are still popular; for example Google's WaveNet, a deep generative model of raw audio waveforms [62], compresses the raw audio using the μ-law algorithm, before further processing by neural networks.

Both μ-law and A-law algorithms perform non-linear dynamic range compression designed to reduce the number of bits of information in each sample of a digital audio signal while preserving the dynamic range of samples at low amplitudes. Fig. 5 shows compression of normalized amplitudes of the input samples, which is approximately linear at low amplitudes and highly non-linear at high amplitudes.

Later, in 1979, perceptual limitations of the human ear were used to encode arbitrary signals [63], which

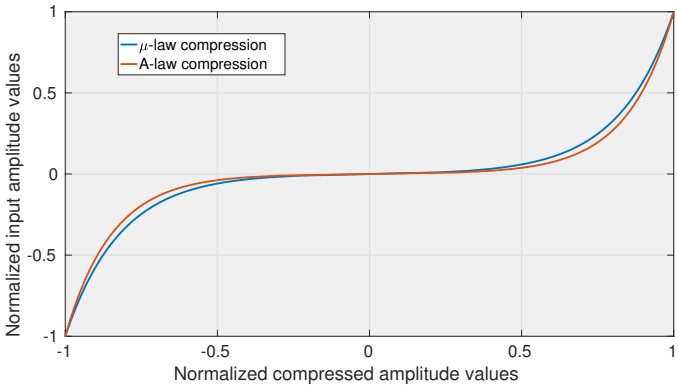

Fig. 5. Comparison of μ-law and A-law algorithms. The μ-law algorithm, primarily used in North America and Japan, provides a slightly larger dynamic range than the A-law, primarily used in Europe. Typically, a 16-bit digital audio signal is reduced to 8-bits by μ-law or A-law encoding.

evolved into perceptual audio coding used nowadays, as exemplified by the MPEG (moving picture experts group) standards. These coders incorporate several psychoacoustic principles, including absolute hearing thresholds, critical band frequency analysis, simultaneous masking, the spread of masking along the basilar membrane, and temporal masking [56]. Imitating the human auditory system, sub-band coding breaks the signal into a number of different frequency bands [64]. This coding indeed resembles spectrotemporal feature organization of A1 discussed in the previous section. Sub-band adaptive differential pulse code modulation with a bit rate of 64 kbit/s is standardized as the G.722 codec. It is also a key component of the popular MP3 format. Perceptual coding is also termed open-loop, since there is no feedback from the output to the input.

In addition, perceptual speech quality assessment is an established area of speech coding aiming at automatically (and non-intrusively, without a reference) evaluating the quality of transmitted speech. For example, the recent Perceptual Objective Listening Quality Assessment POLQA method [65], standardized as ITU-T Rec. P.863, includes a perceptual model based on both spectral and temporal masking effects of human hearing, and cognitive modelling. While the perceptual model based on a gammatone filterbank determines which distortions can be perceived by listeners, the cognitive model predicts the level of those distortions. The output of the cognitive model is the absolute category rating, the overall quality score, which reflects the opinion of an average listener who is used to using commercial telephony services.

B. Production-Model Based Compression

Perceptual audio coding uses models of human auditory perception, whereas speech coding is traditionally based on the human vocal tract (speech production) model. Though the band-limited wired and wireless communication systems have changed dramatically from analogue to digital, the paradigm of speech coding has remained the same, based on the waveform and the linear prediction model [5]. Linear Prediction Coding (LPC) is used in the majority of standardised higher bit-rate [3], [59] and lower bit-rate speech coding [66]–[71].

LPC also uses some models of human auditory perception, such as perceptual weighting of the residual quantization error and adaptive postfiltering [57], in order to minimise different types of auditory distortion. LPC is usually realized as an analysis-by-synthesis system that selects an excitation signal among a large set of candidates in a closed-loop manner. In other words, speech coders include decoded feed-back during encoding. As introduced in Section II-A3, cortical speech processing mechanisms also include feed-back decoding processes. Here the similarity with human cortical speech production ends. Machine speech compression is rather inspired by the physiological process of speech production based on the source-filter theory.

Further connections with human production exist in sparse representations that contribute significantly to a low computational complexity. In 1986, sparse excitation was proposed for Code Excited Linear Prediction (CELP) coding [72] as a complexity reduction method; the speech source defined as a codebook populated with pseudo-random white sequences (Gaussian excitation vectors) was sparse, in terms of the number of nonzero pulses for voiced speech. In a typical 5-ms frame (or sub-frame, depending on the CELP variant) period, only about 10% of the pulses were set to a value other than zero. Laflamme et al. in 1990 introduced sparse algebraic codes (with few nonzero components) for fast searching of the codebooks [73] to get a minimum variance residual with an analysis-by-synthesis scheme. Most current standardized speech coders are based on this Algebraic-CELP coding (ACELP).

Alternatively, a sparse LP residual can be defined within a compressive sampling framework [74], introduced in Section IV.

C. Switched Audio/Speech Coding

Current hybrid coding approaches offer real-time switching between perceptual coding for audio sources, and ACELP coding for speech sources. Two switching coders are nowadays popular: i) Opus [75], the open

source codec of the Internet Engineering Task Force that includes speech coding technology from Skype's SILK codec and audio coding technology from CELT codec (http://celt-codec.org.), and ii) EVS [3], the Enhanced Voice Services codec of the 3rd Generation Partnership Project (3GPP). Recent subjective evaluation showed that the 3GPP EVS codec, compared to Opus, provides the same quality at about half the bitrate in low bitrates [60].

Although recent speech codecs offer new features and improvements for low-delay real-time communication systems, higher quality for clean/noisy speech, mixed content and music, including support for wideband, super-wideband and full-band content, the core compression method is ACELP proposed in 1987 by Adoul et al. [4]. The compression paradigm thus did not change significantly in the last 30 years.

IV. Machine Cognitive Speech Coding (CSC)

Human cognitive speech processing involves transforming sensory inputs in both *feed-forward (bottom-up)* and *feed-back (top-down)* processes. LPC also involves *open-loop (feed-forward)* processing for calculation of the gross spectral shape $1/A(z)$ and *closed-loop (feed-back)* processing for calculation of the excitation signal $U(z)$ that models the fine spectral structures, as the LP model represents the speech signal $S(z)$ as a linear time-invariant system with the following transfer function:

$$S(z) = \frac{U(z)}{A(z)} = \frac{U(z)}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{1}$$

The input to the system is $U(z)$, and p pairs of complex-conjugate poles of $A(z)$ represent formant frequencies (the spectral peaks of the sound spectrum).

Table I compares LPC and Cognitive Speech Coding (CSC). The main differences lie in the underlying model used: LPC is a linear time-invariant system with a pre-defined transfer function and codebooks whose parameters are estimated from the input signal, whereas CSC is a neural network system with learnt parameters using a machine learning approach. Concerning temporal context, LPC has a synchronous (fixed) frame-subframe structure, a wider frame of $20\,\mathrm{ms}$ for calculation of the p LPC coefficients and a narrow frame of $5\,\mathrm{ms}$ for estimation of the LPC excitation signal u. On the contrary, speech communication is known to be an asynchronous process due to asynchronous evolution of various articulatory feature streams [16].

A. Target Speech Representation

At present, speech recognition and synthesis are highly dependent on machine learning tools and big data, and speech coding can benefit from it, for example by learning of better speech representation [76]. However, there is almost no utilization of the overall code (underlying structure) of spoken language. The notion of a code implies relations between message units and signal units [77]. It is known that articulatory interpretation of auditory spectrograms is a key to its understanding, however it becomes more elusive when applied to brain function. We can hypothesize that once the speech code is deciphered, we could design very effective speech compression algorithms, approaching the efficiency of cognitive processing that operates at about 50 bits/second [1].

The speech code representation is usually studied in neurolinguistics [78], [79] without necessary technology transfer to communication engineering. The motor theory of speech perception [80] and the direct realist theory of speech perception [79] claim that the same set of invariants is shared in speech perception and production. The existence of invariant speech representation is greatly debated in motor control, psycholinguistics, neuropsychology and speech neuroscience. Recent findings suggest that the representation is based on auditory and somatosensory speech production parameters [81], and known time-varying articulatory gestures [26], [33], [82], [83].

There is currently no analytic solution for the speech code representation. However, neuropsychological and brain imaging work indicates that language learning produces dedicated neural networks that code the patterns of native-language speech [84]. This also led speech engineers to investigate neural-network based speech coding.

B. Neural Network Based Speech Coding

Li Deng and others have demonstrated that deep auto-encoders can discover some good, discrete representations or "codes" for the entire speech spectrum [85]. The proposed auto-encoder was designed as a deep, five-layer network, with a middle coding layer, where the real-valued activations of hidden units are quantized to be either zero or one with 0.5 as the threshold. These binary codes are then used to reconstruct the original spectrogram. The authors showed improvements over a conventional vector quantization coder with the Linde-Buzo-Gray algorithm. The binary nature of the code resembles the binary nature of phonological speech representation, which is believed to be key in organization of the speech sounds in human brains [25].

Phonological features lie on low-dimensional subspaces. These low-dimensions pertain to either *physiological structures* of the speech production mechanism

TABLE I
COMPARISON OF LPC AND COGNITIVE SPEECH CODING (CSC).

Condition	Linear predictive coding	Cognitive speech coding
Speech representation	formants and vocal tract excitation	articulatory, auditory and somatosensory targets [38]
Models	electrical circuits and digital filters	deep and spiking neural networks
Temporal context	synchronous frame-subframe structure	asynchronous streams [13]
Sparsity	excitation signal [73]	whole speech representation

or the *linguistic structures* of the supra-segmental information. At the physiology level, only certain (very few) combinations of the phonological features can be realized through human vocalization. This property can be formalized by constructing a codebook of structured sparse codes for phonological feature representation. Likewise, at the linguistic level, only some (very few) supra-segmental (e.g. syllabic) mapping of the sequence of phonological features is linguistically permissible. This property can be exploited for block-wise coding of these features with a slower (supra-segmental) dynamic.

1) Short-term (physiological) coding: Figure 6 shows a well-known channel speech coding scheme. Sub-band coding splits the signal into different frequency bands, imitating the human auditory system. Similarly, sub-phonetic coding splits the signal into different phonological classes, imitating phonological processing of the central auditory system. Each phonological class leaves an acoustic signature that listeners can track, similarly as shown in [79]. Parallel feature transmission facilitates asynchronous streams evolution.

Phonological analysis starts with a short-term analysis of speech, which consists of converting the speech signal into a sequence of acoustic feature vectors $X = \{\vec{x}_1, \ldots, \vec{x}_n, \ldots, \vec{x}_N\}$. Each $\vec{x}_n$ is also known as an acoustic frame or just frame, and can be composed by the conventional Mel frequency cepstral coefficients (MFCC). The Mel scale is a perceptual scale often used in speech signal processing. N is the number of frames and the frames are equally spaced in time.

Then, K phonological probabilities z_n^k are estimated for each frame. Each probability is computed independently using a binary classifier based on deep neural network (DNN) and trained with one phonological class versus the rest. Finally, the acoustic feature observation sequence X is transformed into a sequence of phonological vectors $Z = \{\vec{z}_1, \ldots, \vec{z}_n, \ldots, \vec{z}_N\}$. Each vector $\vec{z}_n = [z_n^1, \ldots, z_n^k, \ldots, z_n^K]^\top$ consists of phonological class posterior probabilities $z_n^k = p(c_k|x_n)$ of K phonological features (classes) c_k. The a posteriori estimates $p(c_k|x_n)$ are $0 \leq p(c_k|x_n) \leq 1, \forall k$, and $\max \sum_{k=1}^{K} p(c_k|x_n) = K$. The z_n^k features can be further quantized or compressed using sparse coding that relies on the structured sparsity of the phonological features.

Asaei et al. [53] demonstrated that structured sparse coding of the binary features enables the codec to operate at $700\,\mathrm{bps}$ without imposing any latency or quality loss with respect to the earlier developed vocoder [55]. By considering a latency of about $256\,\mathrm{ms}$, the bit rate of about $300\,\mathrm{bps}$ is achieved without requirement for any prior knowledge on supra-segmental (e.g. syllabic) identities.

Compressive sampling relies on sparse representation to reconstruct a high-dimensional data using very few linear non-adaptive observations. A data representation $\alpha \in \mathbb{R}^N$ is K-sparse if only $K \ll N$ entries of α have nonzero values. We call the set of indices corresponding to the non-zero entries as the support of α. The CS theory asserts that only $M = O(K \log(N/K))$ linear measurements, $z \in \mathbb{R}^M$ obtained as

$$z = D\alpha \qquad \text{(CS coder)} \quad (2)$$

suffice to reconstruct α, where $D \in \mathbb{R}^{M \times N}$ is a *compressive measurement matrix* which preserves the pairwise distances of the sparse features α in the compressed code z.

At the coding step, the choice of *compressive measurement matrix* D is very important. A sufficient but not necessary condition on D to guarantee decoding of the sparse representation coefficients is that all pairwise distances between K-sparse representations must be well preserved in the observation space or equivalently all subsets of K columns taken from the measurement matrix are nearly orthogonal. This condition on the compressive measurement matrix is referred to as the restricted isometry property (RIP). The random matrices generated by sampling from Gaussian or Bernoulli distributions are proved to satisfy RIP condition [87]. It was shown that a choice of Bernoulli matrix is demonstrated to achieve higher robustness to quantization [53].

Given an observation vector z, and the measurement matrix D, the sparse representation α is obtained by the optimization problem stated as

$$\min_{\alpha} \parallel \alpha \parallel_0 \quad \text{subject to} \quad z = D\,\alpha \qquad \text{(Sparse decoder)} \quad (3)$$

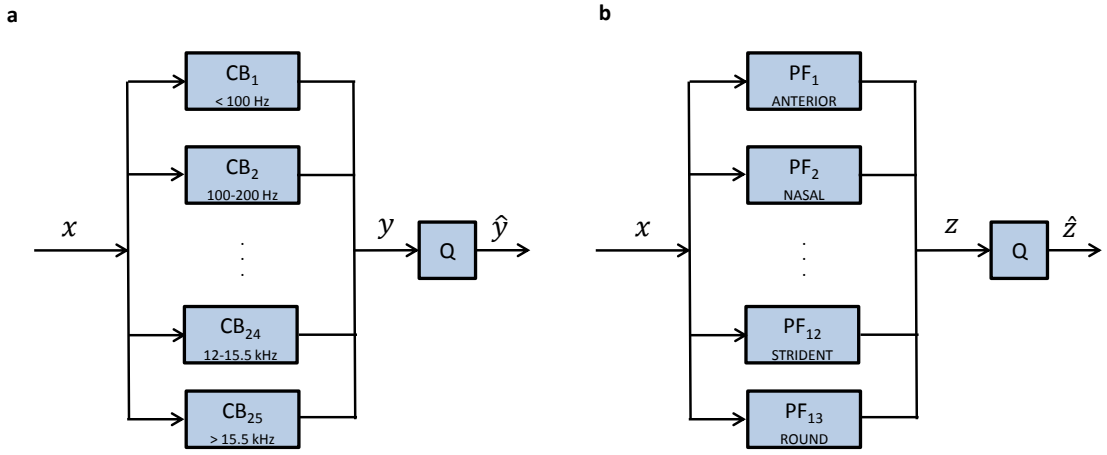

Fig. 6. Channel speech coding. **a)** Channels in sub-band coding [64] are composed of the critical band filters, inspired by Homer Dudley's channel vocoding [86] from 1939. **b)** Channels in sub-phonetic coding [55] are composed of the neural network based phonological filters Deep Neural Networks (DNNs). Phonological features are known also as distinctive or phone-attribute features.

where the counting function $\| \ . \ \|_0$: $\mathbb{R}^M \longrightarrow \mathbb{N}$ returns the number of non-zero components in its argument. The non-convex objective $\| \ \alpha \ \|_0$ is often relaxed to $\| \ \alpha \ \|_1 = \sum_i |\alpha_i|$ which can be solved in polynomial time [88]. Recent advances in CS exploits inter-dependency structure underlying the support of the sparse coefficients in recovery algorithms to reduce the number of required observations and to better differentiate the true coefficients from recovery artifacts for higher quality [89].

2) Long-term (linguistic) coding: Long-term analysis includes analysis of speech information encoded at "stress" δ (1–3 Hz) and "syllabic" θ (4–8 Hz) timescales. An important module that facilitates extraction of this information is robust syllable boundary detector.

The neuromorphic syllable detector [51] has some 'desired' properties that are very suitable for speech processing systems. Similarly as humans encode speech incrementally, i.e., not considering future temporal context, the proposed method works incrementally as well. In addition, it is highly robust to noise. Syllabification performance at different noise conditions was compared to the existing Mermelstein and group delay algorithms. While the performance of the existing methods depend on the type of noise and signal to noise ratio, the performance of the proposed method is constant under all noise conditions.

Figure 7 shows the neural oscillation that automatically lock to speech slow fluctuations that convey the syllabic rhythm.

3) Composition of short and long-term neural networks: A Neural Network (NN) based speech coder can be realized as a composition of deep and spiking neural networks [90]. The deep neural networks encode and decode the speech signal based on the binary phonological speech representation, and the spiking net based on neuromorphic syllabification, described above, is used for prosody encoding. Prosody represents the patterns of stress and intonation of the speech signal. Figure 8 shows the training and inference stages of the three different NNs.

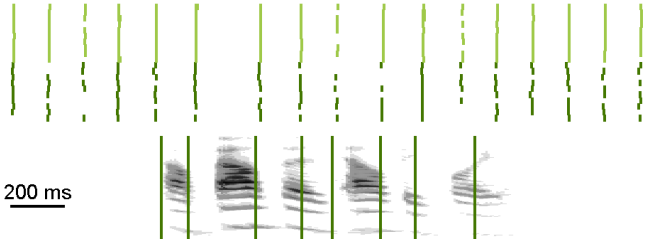

Fig. 7. *Neuromorphic model output for one examplar sentence ("Alfafa is healthy for you"). Dark green ticks on top represent excitatory neurons spikes, light green ticks represent inhibitory neuron spikes. Vertical lines on top of spectrogram represent actual syllable boundaries.*

Decoding is realized as a synthesis DNN that learns the highly-complex mapping of the transmitted sequence, Z, to the speech parameters. It consists of two computational steps. The first step is a DNN forward pass that generates the speech parameters, and the second one is generation of the speech samples from the speech parameters including a decoded pitch signal. The pitch signal is coded with a codebook that contains the logarithm of the continuous pitch of all the syllables of the training data, parametrized with the discrete Legendre orthogonal polynomials.

Intelligibility evaluation of this NN codec showed about 10% degradation, when comparing with LPC (Speex) coding; however, it operates at a bit rate ap-

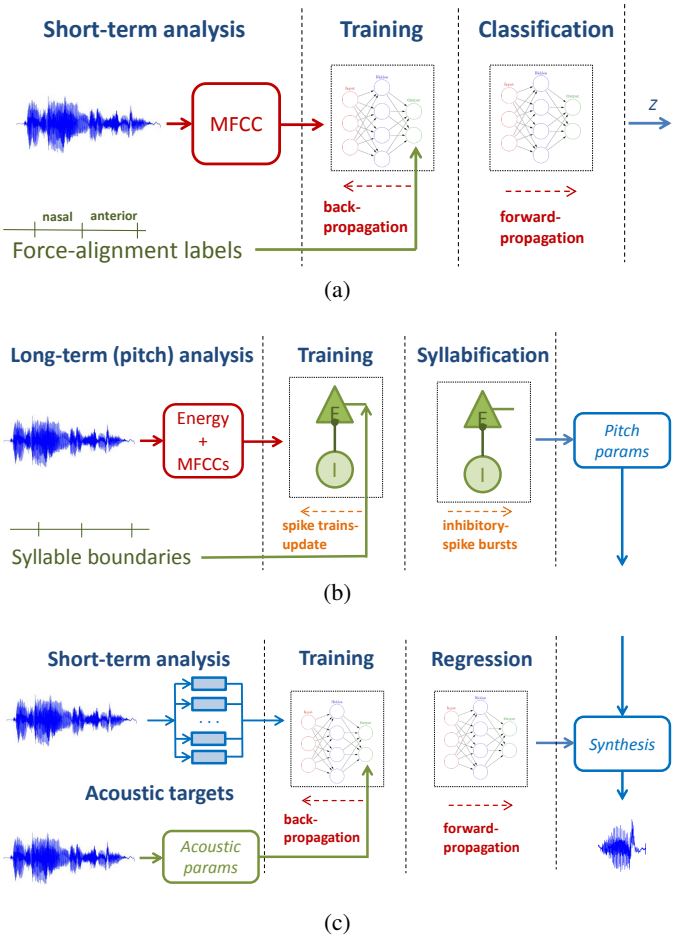

Fig. 8. Training and inference stages for an analysis DNN shown on 8a, a spiking NN shown on 8b, and a synthesis DNN shown on 8c. Training of the analysis DNN and the spiking NN requires phonetic and syllabic boundary labels, respectively, whereas training of the synthesis DNN does not require force-aligned labels. Pitch is the fundamental frequency of the sound signal.

proximately 6 times lower.

C. Challenges

This sections outlines current challenges in cognitive speech coding:

1) CSC Model Architecture: What type, size and composition of neural networks are suitable for CSC?

It was shown that more than 77% of all coding distortion of NN codec comes from the parametric vocoding used for speech re-synthesis. This challenge can be thus inspired by recent advancements in end-to-end speech synthesis research.

Modelling of the long-term context in CSC can be investigated by adapting spiking or recurrent neural networks.

2) CSC Speech Parameters: What is an efficient speech representation for CSC, and how to design its sparse coding?

CSC speech parameters should be based on cortical representation introduced in Section II. It includes sparse short-term (physiological) and long-term (linguistic) representations. The parameters have to be efficiently estimated by the CSC model.

3) CSC Adaptability: How to realize computational sensory feed-back?

The goal is to implement an adaptation of the cortical speech representation based on sensory feed-back. 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. CSC introduces a novel feedback for the speech code constructed from the auditory and articulatory targets, currently not used by waveform/LPC coding.

An example is shown for adaptive speech activity detection [44]. The system employs a 2-D Gabor filter bank whose parameters are retuned offline to improve the separability between the feature representation of speech and nonspeech sounds, and it attempts to minimize the misclassification risk of mismatched data, with respect to the original statistical models.

4) CSC Robustness: How to focus an attention of CSC to increase intelligibility or to decrease cognitive load?

Cognitive speech coding can benefit from biological cognitive aspects of speech communication, trying to design speech codecs that "are aware" of the hidden structure of the transmitted speech signal (speech code), and the transmitted code is linguistically relevant. Linguistically relevant transmission code could bring novel functionality to speech transmission systems, performing tasks such as automatic dialect correction and pronunciation improvements for people with phonological and articulatory disorders that might eventually lead to intelligibility enhancements.

Error minimization can perform error correction as defined by the DIVA and HSFC models. For example, let us consider the task of dialect correction. Error minimization might be implemented as an automatic accentedness evaluation of non-native speech [91], [92] in an closed-loop fashion.

Cognitive load, related to the cost of cognitive processing resources and listening effort, is at present studied in the field of cognitive hearing science and by manufacturers of hearing aids devices. CSC can be inspired by this ongoing research to apply their results into intelligibility enhancements of transmitted speech signals.

5) CSC Language Independence: How to make neural-network based CSC language independent?

Language independent speech and language technology belongs to the main research topics of the speech signal processing community. One approach is in modeling all the International Phonetic Alphabet (IPA) symbols. There are more than 100 the IPA symbols that might be further extended with more narrow phonetic and prosody annotations. On the contrary, all the speech sounds share some attributes, known in linguistics as phonological features or phone-attributes in the speech community. In phonology, one example could be the Sound Pattern of English set consisting of 13 features [93]. In the speech community, one major effort resulted into Automatic Speech Attribute Transcription [94], which proposes a framework for speech understanding based on processing of cognitive hypotheses created from the acoustic and auditory cues of the phone-attributes. Recent phonetic DNN training analysis confirmed that the hidden layers learn an effective representational basis, the phone-attributes, for the formation of invariant phonemic categories [95]. Exploring this phonetic invariance across the languages is thus a promising direction also for the multi-lingual CSC technology.

6) CSC speaker recognizability: How to make neural-network based CSC speaker independent?

Another a very relevant problem is speaker recognizability. Early work on speaker adaptive phonetic vocoding resulted into variable bit-rate coding [96]. Promising results have showed recent advances in NN-based generative models of raw audio, such as WaveNet from Google Research [62] and Deep Voice 2 from Baidu Research [97], capable to learn and imitate hundreds of voices within a single generative model.

7) CSC Complexity: How to minimize computational complexity of CSC?

Complexity of compressive sampling is about $O(K\log(N/K))$ for K-sparse ($K \ll N$) N-dimensional speech representation.

Computational complexity of a DNN is about N_w, where w is the number of weights of the DNN. The compositional neural network based codec described above consisted of 12 analysis DNN, each trained with 2.46 million parameters, and one synthesis DNN trained with 3.31 million parameters. Decoding with generative models of raw audio with based on the advanced NN architectures, composed of deep convolutional and recurrent NNs, might have significantly higher computational demand, not feasible for current devices.

V. Summary

We have presented a tutorial of the impact of cognitive speech processing on speech compression. We described basic concepts of human cognitive speech processing, and how cognitive speech coding has impacted current audio and speech coding systems. Properties of cognitive speech coding were then presented, and compared to linear predictive coding.

Cognitive speech coding has great potential to further impact current speech coding standards. Deep learning and neural basis of speech and language processing have been tremendously advanced recently, and many of the findings can be transferred to the field of speech coding. The tutorial ends by outlining current challenges in CSC, relying on incorporating phonological posteriors and temporal models, employing speech production models and cognitive feedback for enhanced speech compression.

References

[1] M. S. Lewicki, "Efficient coding of natural sounds," *Nature neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[2] A. S. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[3] 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*. IEEE, Apr. 2015, pp. 5698–5702.

[4] J.-P. Adoul, P. Mabilleau, M. Delprat, and S. Morissette, "Fast celp coding based on algebraic codes," in *Proc. of ICASSP*, vol. 12. IEEE, 1987, pp. 1957–1960.

[5] B. S. Atal, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 154–161, Mar. 2006.

[6] P. Vaidyanathan, "The theory of linear prediction," *Synthesis lectures on signal processing*, vol. 2, no. 1, pp. 1–184, 2007.

[7] T. Chi, P. Ru, and S. a. Shamma, "Multiresolution spectrotemporal analysis of complex sounds." *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, aug 2005.

[8] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393–402, May 2007.

[9] D. Pisoni and R. Remez, *The handbook of speech perception*. John Wiley & Sons, 2008.

[10] A.-L. L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations." *Nature neuroscience*, vol. 15, no. 4, pp. 511–517, Apr. 2012.

[11] R. L. Diehl, A. J. Lotto, and L. L. Holt, "Speech perception," *Annu. Rev. Psychol.*, vol. 55, pp. 149–179, 2004.

[12] D. Poeppel, "The neuroanatomic and neurophysiological infrastructure for speech and language," *Current Opinion in Neurobiology*, vol. 28, pp. 142–149, Oct. 2014.

[13] M. Cernak, P. N. Garner, A. Lazaridis, P. Motlicek, and X. Na, "Incremental Syllable-Context Phonetic Vocoding," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1019–1030, Jun. 2015.

[14] S. Zeki, "A massively asynchronous, parallel brain," *Phil. Trans. R. Soc. B*, vol. 370, no. 1668, p. 20140174, 2015.

[15] R. A. Stevenson, N. A. Altieri, S. Kim, D. B. Pisoni, and T. W. James, "Neural processing of asynchronous audiovisual speech perception," *Neuroimage*, vol. 49, no. 4, pp. 3308–3318, 2010.

[16] R. Rasipuram and M. Magimai-Doss, "Articulatory feature based continuous speech recognition using probabilistic lexical modeling," *Computer Speech & Language*, vol. 36, pp. 233–259, 2016.

[17] A. D. Friederici, "The cortical language circuit: from auditory perception to sentence comprehension," *Trends in Cognitive Sciences*, vol. 16, no. 5, pp. 262–268, apr 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1364661312000794

[18] Y. Grodzinsky and A. Santi, "The battle for Broca's region," *Trends in Cognitive Sciences*, vol. 12, no. 12, pp. 474–480, dec 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661308002222?via{\%}3Dihub

[19] E. Smith and M. S. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computation*, vol. 17, pp. 19–45, 2005.

[20] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[21] ——, "Sparse coding of sensory inputs," *Current opinion in neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[22] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean *et al.*, "On rectified linear units for speech processing," in *Proc. of ICASSP*. IEEE, 2013, pp. 3517–3521.

[23] T. Hromádka and A. M. Zador, "Representations in auditory cortex," *Current opinion in neurobiology*, vol. 19, no. 4, pp. 430–433, 2009.

[24] A.-L. Giraud and D. Poeppel, "Speech perception from a neurophysiological perspective," in *The human auditory cortex*. Springer, 2012, pp. 225–260.

[25] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic Feature Encoding in Human Superior Temporal Gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, Feb. 2014.

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

[27] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Phoneme representation and classification in primary auditory cortex," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 899–909, 2008.

[28] A. Gafos and L. Goldstein, "Articulatory representation and phonological organization," in *The Oxford Handbook of Laboratory Phonology*, A. C. Cohn, C. Fougeron, and M. K. Huffman, Eds. Oxford University Press, 2012.

[29] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, p. 303, 1995.

[30] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility." *PLoS computational biology*, vol. 5, no. 3, p. e1000302, mar 2009.

[31] J. Gross, H. Nienke, G. Thut, P. Schyns, S. Panzeri, P. Belin, and S. Garrod, "Speech rhythms and multiplexed oscillatory sensory coding in the human brain." *PLoS biology*, vol. 11, no. 12, p. e1001752, dec 2013.

[32] D. Poeppel, "The Analysis of Speech in Different Temporal Integration Windows: Cerebral Lateralization As 'Asymmetric Sampling in Time'," *Speech Communication*, vol. 41, no. 1, pp. 245–255, Aug. 2003.

[33] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation." *Nature*, vol. 495, no. 7441, pp. 327–332, Mar. 2013.

[34] P. Gagnepain, R. N. Henson, and M. H. Davis, "Temporal predictive codes for spoken words in auditory cortex." *Current biology : CB*, vol. 22, no. 7, pp. 615–21, apr 2012.

[35] J. L. McClelland, D. Mirman, D. J. Bolger, and P. Khaitan, "Interactive Activation and Mutual Constraint Satisfaction in Perception and Cognition." *Cognitive science*, pp. 1–51, aug 2014.

[36] D. Norris and J. M. McQueen, "Shortlist B: a Bayesian model of continuous speech recognition." *Psychological review*, vol. 115, no. 2, pp. 357–95, apr 2008.

[37] K. J. Friston, "A theory of cortical responses." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 360, no. 1456, pp. 815–36, apr 2005.

[38] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and Cognitive Processes*, vol. 26, no. 7, pp. 952–981, Aug. 2011.

[39] G. Hickok, "The architecture of speech production and the role of the phoneme in speech processing," *Language, Cognition and Neuroscience*, vol. 29, no. 1, pp. 2–20, Jan. 2014.

[40] S. Sundaram and S. Narayanan, "Discriminating two types of noise sources using cortical representation and dimension reduction technique," in *Proc. of ICASSP*, vol. 1. IEEE, 2007, pp. I–213.

[41] W. Jeon and B.-H. Juang, "Speech analysis in a model of the central auditory system," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1802–1817, 2007.

[42] A. Hyafil and M. Cernak, "Neuromorphic Based Oscillatory Device for Incremental Syllable Boundary Detection," in *Proc. of Interspeech*, Sep. 2015, pp. 1191–1195.

[43] C. Spille, B. Kollmeier, and B. T. Meyer, "Combining binaural and cortical features for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 756–767, 2017.

[44] A. Bellur and M. Elhilali, "Feedback-driven sensory mapping adaptation for robust speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 481–492, 2017.

[45] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[46] R. Ullmann and H. Bourlard, "Predicting the intrusiveness of noise through sparse coding with auditory kernels," *Speech Communication*, vol. 76, pp. 186–200, 2016.

[47] N. L. Carlson, V. L. Ming, and M. R. DeWeese, "Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus," *PLoS Comput Biol*, vol. 8, no. 7, p. e1002594, 2012.

[48] S. V. David, N. Mesgarani, and S. A. Shamma, "Estimating sparse spectro-temporal receptive fields with natural stimuli," *Network: Computation in Neural Systems*, vol. 18, no. 3, pp. 191–212, 2007.

[49] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 416–422, 2002.

[50] N. Mesgarani and S. Shamma, "Speech processing with a cortical representation of audio," in *Proc. of ICASSP*. IEEE, 2011, pp. 5872–5875.

[51] A. Hyafil, L. Fontolan, C. Kabdebon, B. Gutkin, A.-L. Giraud, and H. Brownell, "Speech encoding by coupled cortical theta and gamma oscillations," *eLife*, May 2015.

[52] 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.

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

[54] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, pp. 333–382, 1989.

[55] M. Cernak, B. Potard, and P. N. Garner, "Phonological vocoding using artificial neural networks," in *Proc. of ICASSP*. IEEE, Apr. 2015, pp. 4844–4848.

[56] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

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

[58] J. Herre and M. Lutzky, "Perceptual audio coding of speech signals," in *Springer Handbook of Speech Processing*, J. Benesty, Sondhi, and Y. Huang, Eds. Springer Berlin Heidelberg, 2008, pp. 393–410.

[59] J. Gibson, "Speech Compression," *Information*, vol. 7, no. 2, pp. 32+, Jun. 2016.

[60] A. Rämö and H. Toukomaa, "Subjective quality evaluation of the 3gpp evs codec," in *Proc. of ICASSP*. IEEE, 2015, pp. 5157–5161.

[61] J. B. Allen, "Nonlinear cochlear signal processing and masking in speech perception," in *Springer Handbook of Speech Processing*, J. Benesty, Sondhi, and Y. Huang, Eds. Springer Berlin Heidelberg, 2008, pp. 27–60.

[62] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[63] M. R. Schroeder, B. S. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[64] M. Vetterli and J. Kovačevic, *Wavelets and Subband Coding*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1995.

[65] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part itemporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[66] 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, no. 5, pp. 770–780, Oct. 1982.

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

[68] ——, "A segment vocoder at 150 b/s," in *Proc. of ICASSP*, vol. 8. IEEE, Apr. 1983, pp. 61–64.

[69] 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. IEEE, Apr. 1983, pp. 65–68.

[70] 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, no. 3, pp. 537–545, Jun. 1985.

[71] Y. Shiraki and M. Honda, "LPC speech coding based on variable-length segment quantization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 36, no. 9, pp. 1437–1444, Sep. 1988.

[72] G. Davidson and A. Gersho, "Complexity reduction methods for vector excitation coding," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, vol. 11. IEEE, 1986, pp. 3055–3058.

[73] C. Laflamme, J.-P. Adoul, H. Su, and S. Morissette, "On reducing computational complexity of codebook search in celp coder through the use of algebraic codes," in *Proc. of ICASSP*. IEEE, 1990, pp. 177–180.

[74] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction," *IEEE Signal processing letters*, vol. 17, no. 1, pp. 103–106, 2010.

[75] J.-M. Valin, K. Vos, , and T. T.B., "Definition of the Opus audio codec," Sept. 2012. [Online]. Available: http://tools.ietf.org/html/rfc6716

[76] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *Proc. of ICASSP*. IEEE, 2011, pp. 5884–5887.

[77] G. Fant, "On the speech code," KTH Computer Science and Communication, Technical Report, 2001.

[78] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code." *Psychological review*, vol. 74, no. 6, pp. 431–461, Nov. 1967.

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

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

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

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

[83] ——, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[84] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831–843, 2004.

[85] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder." in *Proc. of Interspeech*. Citeseer, 2010, pp. 1692–1695.

[86] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[87] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21–30, 2008.

[88] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab software for disciplined convex programming," 2008.

[89] A. Asaei, H. Bourlard, and V. Cevher, "Model-based compressive sensing for multi-party distant speech recognition," in *Proc. of ICASSP*, 2011.

[90] 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, no. 12, pp. 2301–2312, Dec 2016.

[91] 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*, 2015, pp. 648–652.

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

[93] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, NY: Harper & Row, 1968.

[94] C.-H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, and L. R. Rabiner, "An overview on automatic speech attribute transcription (asat)," in *Proc. of Interspeech*, 2007, pp. 1825–1828.

[95] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Proc. of Interspeech*, 2015, pp. 1912–1916.

[96] C. M. Ribeiro and I. Trancoso, "Phonetic vocoding with speaker adaptation." in *EUROSPEECH*, 1997, pp. 1291–1294.

[97] S. O. Ark, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep Voice 2: Multi-Speaker Neural Text-to-Speech," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. [Online]. Available: http://research.baidu.com/wp-content/uploads/2017/05/Deep-Voice-2-Complete-Arxiv.pdf