Composition of Deep and Spiking Neural Networks for Very Low Bit Rate Speech Coding

Milos Cernak, *Member, IEEE,* Alexandros Lazaridis, *Member, IEEE,* Afsaneh Asaei, *Senior Member, IEEE,*
Philip N. Garner, *Senior Member, IEEE*

***Abstract*—Most current very low bit rate (VLBR) speech coding systems use hidden Markov model (HMM) based speech recognition and synthesis techniques. This allows transmission of information (such as phonemes) segment by segment; this decreases the bit rate. However, an encoder based on a phoneme speech recognition may create bursts of segmental errors; these would be further propagated to any suprasegmental (such as syllable) information coding. Together with the errors of voicing detection in pitch parametrization, HMM-based speech coding leads to speech discontinuities and unnatural speech sound artefacts.**

In this paper, we propose a novel VLBR speech coding framework based on neural networks (NNs) for end-to-end speech analysis and synthesis without HMMs. The speech coding framework relies on a phonological (sub-phonetic) representation of speech. It is designed as a composition of deep and spiking NNs: a bank of phonological analysers at the transmitter, and a phonological synthesizer at the receiver. These are both realised as deep NNs, along with a spiking NN as an incremental and robust encoder of syllable boundaries for coding of continuous fundamental frequency (F0). A combination of phonological features defines much more sound patterns than phonetic features defined by HMM-based speech coders; this finer analysis/synthesis code contributes to smoother encoded speech. Listeners significantly prefer the NN-based approach due to fewer discontinuities and speech artefacts of the encoded speech. A single forward pass is required during the speech encoding and decoding. The proposed VLBR speech coding operates at a bit rate of approximately 360 bits/s.

***Index Terms*—Very low bit rate speech coding, deep neural networks, spiking neural networks, continuous F0 coding**

I. Introduction

The ITU-T standardisation effort for speech coders operating below 4000 bits per second (bit/s) began in 1994 [1]. However, it has been shown to be difficult to achieve toll-quality performance in all conditions, such as intelligibility, quality, speaker recognizability, communicability, language independence and complexity. All these conditions are exposed even more in speech coders operating at VLBR speech coding of the order of hundreds of bit/s.

To achieve VLBR of 200 to 500 bit/s, parametric speech coding based on a recognition/synthesis paradigm has been proposed. Approaches following this paradigm can be classified into two categories: corpus-based, e.g., [2], [3], and hidden Markov model (HMM) based, e.g., [4]–[6]. This classification follows trends in speech synthesis, where popular unit-selection methods are replaced by HMM-based parametric speech synthesis methods. The parametric methods benefit from better adaptation properties and lower footprint. However, most current HMM-based VLBR systems have complex designs. The phonetic encoder — automatic speech recognition (ASR) — consists of acoustic HMMs and language models and an incremental search module. Similarly, the phonetic decoder requires acoustic HMMs, including a streaming/performative HMM-based speech synthesis system and an incremental speech vocoder.

Our recent work [7] also focused on HMM-based VLBR speech coding, designing a coder operating with an acceptable communication delay for real-time speech communication, and with a view to be exploited in military and tactical communication systems.

Closer analysis of our HMM-based encoder — a phoneme ASR system — revealed that it sometimes created bursts of segmental errors when the recognition failed. This is a well known phenomenon; phoneme ASR misrecognition rates tend to increase with longer words [8]. In addition, the coder [7] detected syllable boundaries from recognised phoneme sequences based on the sonority sequencing principle; thus phoneme misrecognition was further propagated to the syllable boundary estimation sometimes resulting in wrongly detected syllables. The HMM coder also used voiced/unvoiced detection for parametrization of the F0 signal in the voiced regions. This F0 encoding, plugged-in to the VLBR system, created additional speech discontinuities and unnatural speech sound artefacts. Moreover, canonical phone indexes transmitted from a transmitter to a receiver compressed all phonetic variability to the order of tens of phonetic categories. Increasing the number of categories increased the bit rate.

In this paper, we aim to address the above limitations of such speech codecs for very low bit rates that are composed of HMM-based recognition/synthesis components. First, we propose to *replace HMMs by deep NNs* that perform speech analysis and synthesis based on a phonological speech representation. The phonological speech representation extends the encoded phonetic variability to the order of hundreds or even thousands of sound categories. Second, we propose to use *a spiking NN* as a neuromorphic, incremental, and highly noise-robust syllable boundary detector, used for syllable-based continuous F0 signal coding. Using continuous F0 modelling should also alleviate speech re-synthesis discontinuities caused by erroneous detection of unvoiced segments. This end-to-end NN based speech coding (called *NN speech coder* hereinafter) should significantly reduce the current design (and hopefully also computational) complexity since the speech encoding and decoding would be realised as a single NN forward pass.

Authors are with Idiap Research Institute, Centre du Parc, Rue Marconi 19, 1920 Martigny, Switzerland, contact details: http://people.idiap.ch/mcernak.

In this work, we consider the three phonological systems defined in Appendix A of [9]: (i) the Government Phonology (GP) [10], [11], (ii) the Sound Pattern of English (SPE) [12] and (iii) the extended SPE system (eSPE) [13], [14]. Each phoneme is represented by its sub-phonetic attributes, or phonological classes. A vector of all phonological class probabilities is referred to as the phonological posterior. There are only very few phonological classes comprising a short term speech signal; hence, the phonological posterior is a sparse vector, allowing very few combinations that can be stored in a codebook for VLBR speech coding [15]. The temporal span of phonological features is wider than the span of phonetic features, and thus the frame shift can be higher, i.e., fewer frames are transmitted yielding lower bit rates.

The rest of the paper is organized as follows. Section II contains a review of usage of NNs for speech coding. In section III an open-source experimental framework used in this work is introduced, and in Section IV the results are presented of a comparison of HMM-based and NN-based VLBR speech coders. Finally, conclusions are drawn in Section VI.

II. Neural Networks for Speech Coding

A. Background

Recently, in the speech recognition/analysis field, a shift has been observed from HMM-based approaches towards the use of deep neural networks (DNNs) [16]. Even though the concept of using neural networks in recognition is not new [17], the increase of available speech resources and of computational power, and the use of graphics processing units, have led to a major research interest in DNNs. Additionally, deep architectures with multiple layers can overcome the limitations in the representational capability of HMMs, which are incapable of modelling multiple interacting source streams [18]. Furthermore, DNNs are able to create non-linear mappings between the input and output features which cannot be achieved by using Gaussian mixture models (GMMs) in HMM-based approaches, making them more appropriate for modelling the speech signal [19]. As a result, DNN-based approaches have shown superior performance in comparison to HMM-based ones in recognition tasks [19], [20].

The same shift has also been observed in the text-to-speech (TTS) field. There are various limitations and drawbacks which occur in HMM-based TTS [21], e.g., inefficiency to express complex dependencies in the feature space [22]; this leads to decision trees becoming exceedingly large, inefficient and data hungry. Also, the use of decision trees leads to the fragmentation of the training data, linking specific parts of them to each terminal node of the tree [23]. These limitations led to the introduction of DNNs in the field of parametric speech synthesis as well, constantly outperforming HMM-based speech synthesis systems [21], [24], [25].

In the context of speech coding, NNs are used in *waveform-approximating coders*, a family of coders originated in [26], [27], to address either improving quality or reducing computational complexity. The former usage aims to replace linear prediction of speech samples or parameters of the excitation signal with a non-linear prediction. The non-linear prediction is usually based on multilayer perceptrons [28]–[32]. Recently, regression-based packet loss concealment was proposed using DNNs [33]. The latter usage aims to reduce the complexity of the codebook search process or gain prediction [34] using, for example, recurrent NNs [35], [36].

Speech coding based on speech modelling is known as *parametric coding*, where the parameters of the speech models are transmitted, as in, e.g., a multiband excitation vocoder [37]. As in waveform coding, a multilayer perceptron was proposed to decrease the computation complexity of the codebook of line spectral frequencies in the 800 $\mathrm{bit/s}$ multiband excitation speech coding [38].

While NNs have been used in previous speech coding approaches only as isolated modules targeting some particular computation, we propose *end-to-end VLBR NN-based coder*, aiming to replace HMM-based speech analysis and synthesis by deep and spiking NNs.

B. Coding of phonetic and phonological information

VLBR coding has its roots in the linear prediction model [39]–[44], as do the majority of standardised higher bit-rate speech codecs [45]. Unlike such coding, which aims to compress the parameters of the linear prediction model, we aim to keep the parameters of the linear model. Rather, we move the compression one level above, from the acoustics to the phonetic and phonological representation of the speech signal. Our previous work [7] showed that, if such a representation is inferred from the acoustics well, we could decrease the transmission rate to about 200 $\mathrm{bit/s}$, however the speech quality did not achieve the quality of old LPC coders operating at 500 $\mathrm{bit/s}$ or less, that is, slightly lower quality than normal LPC coding at 2 $\mathrm{kbit/s}$. In this work, we hypothesise that recent advances in the convergence of deep learning and speech technology could fill the gap, i.e., improving the speech quality of VLBR coding whilst still using a phonetic and phonological speech representation. It is also fair to note that using the phonetic and phonological speech representation is still language dependent, particularly in speech synthesis.

Encoding of segmental information starts with analysis by converting a segment of speech samples into a sequence of acoustic features $X = \{\vec{x}_1, \ldots, \vec{x}_n, \ldots, \vec{x}_N\}$, where N denotes the number of segments in the utterance. Conventional cepstral coefficients can be used as acoustic features. Encoding can be done on a phonetic or phonological (sub-phonetic) level.

In the former case, the acoustic feature observation sequence X is converted into a parameter sequence $\vec{z}_n = [z_n^1, \ldots, z_n^p, \ldots, z_n^P]^\top$, where the n-th frame consists of posterior probabilities $z_n^p = p(c_p|x_n)$ of P classes (phonemes), and $.^\top$ is the transpose operator. The a-posteriori estimates $p(c_p|x_n)$ are $0 \leq p(c_p|x_n) \leq 1, \forall p$ and $\sum_{p=1}^{P} p(c_p|x_n) = 1$. All the phonemes have to be recognised to access higher semantic levels (words and utterances); hence, using the phone posterior probabilities can be considered a sequential scheme. The phonetic vocoding, where only one encoded segment (a phoneme) is transmitted each time, is an example.

In the latter case, the acoustic feature observation sequence X is converted into a parameter sequence $\vec{z}_n = [p(c_1|x_n), \ldots, p(c_k|x_n), \ldots, p(c_K|x_n)]^\top$, that consists of K phonological class-conditional posterior probabilities, where c_k denotes the phonological class. The phonological posteriors are computed by a bank of parallel DNNs, each estimating the posteriors z_n^k as probabilities that the k-th phonological feature occurs (versus does not occur). The a-posteriori estimates $p(c_k|x_n)$ are also $0 \leq p(c_k|x_n) \leq 1, \forall k$, but $\max \sum_{k=1}^{K} p(c_k|x_n) = K$. Only very few classes are active during a short term signal, $\sum_{k=1}^{K} p(c_k|x_n) \ll K$, resulting in a sparse vector $\vec{z}_n$. Using the phonological posterior probabilities can be considered a parallel scheme via K different phonological classes.

While erroneous phone posterior estimation leads to a possible failure of the higher semantic segment recognition, erroneous phonological posterior estimation leads to a failure only at a sub-phonetic feature level, and this partial error does not necessarily lead to misrecognition of the whole recognized segment.

Decoding of segmental information is realised as a DNN that learns the highly-complex mapping of the parameter sequence, Z, to the speech parameters [46]. It consists of two computational steps. The first step is a DNN forward pass that generates the speech parameters (the LPC speech parameters described in Section III-C4); the second one is generation of the speech samples from the speech parameters.

C. *Coding of prosodic information*

Speech analysis results in discrete units while moving from segmental to suprasegmental (prosodic) level of speech representation. This discrete information can be estimated directly from the speech signal. For example, the Probabilistic Amplitude Demodulation method proposed in [47] can robustly estimate the syllable and stress amplitude modulations using a representation of electrophysiological recordings of the auditory cortex. The work of [48] proves that phase relations of the amplitude modulations, known as hierarchical phase locking and nesting, or synchronization across different temporal granularity [49], is a good indication of the syllable stress.

Phonological posteriors have several interesting properties. Even though they are segmental features by definition, they convey prosodic information about lexical stress and prosodic accent; this is embedded in their support (index) of active coefficients [50]. In the context of parametric speech coding it could be interpreted that we do not need to encode this kind of prosodic information explicitly. Rather, in our current approach, we focus just on coding of the continuous F0 signal. Modelling continuous F0 has been shown to be more effective in achieving natural synthesised speech [51], [52], and can be effectively used with noisy speech [53]. We therefore hypothesise that continuous F0 modelling could improve recognition/synthesis VLBR coding as well.

Effective encoding of the F0 signal can be realised by curve fitting done on a syllable level. We thus propose to encode the continuous F0 signal using the discrete (Legendre) orthogonal polynomial (DLOP), as in [7]. To estimate syllable boundaries from the speech signal, a neuromorphic oscillatory device is used, based on modelling brain neural oscillations at syllable frequency. This results in highly noise robust incremental syllable boundary detection [54]. It is built around an interconnected network composed of 10 excitatory and 10 inhibitory leaky integrate-and-fire neurons. The spiking NN declares a putative syllable boundary for each inhibitory spike burst. We selected this approach to alleviate segmental error propagation in the suprasegmental information coding, and also keep an end-to-end neural network design for the VLBR system.

In the original proposal of syllable-based F0 parametrization for speech coding [55], unvoiced syllables were not parametrized (and not transmitted); also the pitch coding operated at a very low 40–60 bit/s. To achieve similar transmission rates with continuous F0 coding, we further linearly quantized the DLOP parameters. Thus, 3-bit quantized second order DLOP (linear) F0 signal stylization is used for coding of the original F0 signal.

D. *Transmission scheme*

Segmental features — phonological posteriors — have values mostly concentrated very close to either 1 or 0, and these binary patterns allow very efficient use of 1-bit quantization: the probabilities above 0.5 are normalized to 1 and the probabilities less than 0.5 are forced to zero.

Figure 1 illustrates a demonstration sample of the transmission scheme. Figure 1a shows the speech signal. Figure 1b shows binary values of the three basic resonance phonological primes of the GP system commonly labelled as A, U, I, denoting the peripheral vowel qualities [a], [u] and [i] respectively. Other vowels are defined by a composition of the basic ones; for example, [e] results from fusing the I and A primes. In addition to these 'vocalic' primes, GP also proposes the "consonantal" primes that are omitted in the picture for simplicity.

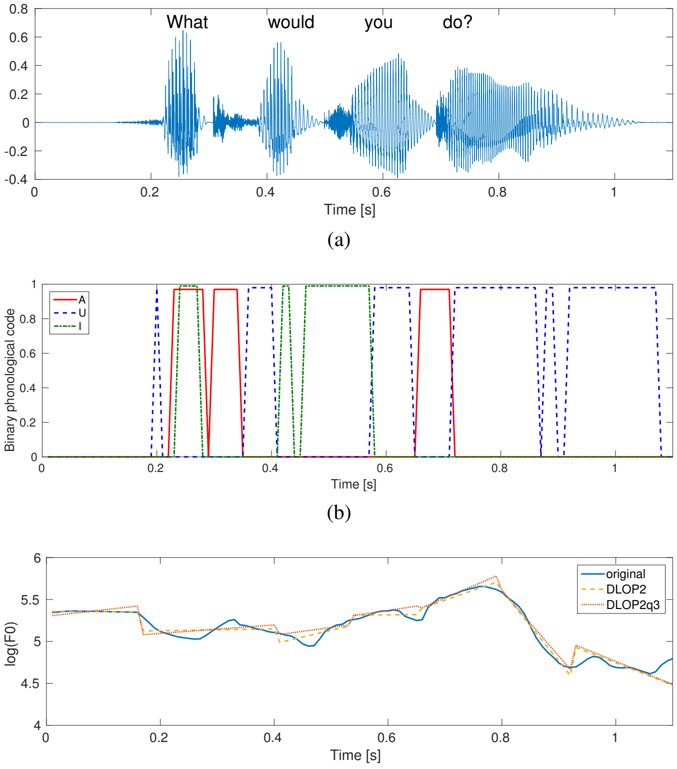
Figure 1c shows an original continuous log F0 signal, and linear curve fitting "DLOP2" using the syllable boundaries from the SNN. Each syllable is parametrized by 2 floating point numbers, that are further quantized using linear 3-bit codebooks, drawn as "DLOP2q3". 6 bits are thus needed to parametrize the first DLOP parameter (the F0 mean) and the second DLOP parameter (the F0 slope) of the transmitted syllable-based prosodic code.

III. Open-Source Experimental Framework

A. *Composition of neural networks*

Figure 2 shows the design of the NN codec. The encoder shown in Figure 2a is based on a bank of DNNs performing segmental speech analysis of conventional acoustic features, and a parallel spiking NN detecting syllable boundaries of the continuous F0 signal. The decoder shown in Figure 2b is based on a DNN performing synthesis of speech cepstral parameters from transmitted segmental and prosodic information.

The outputs of segmental speech analysis are phonological posteriors where all unique patterns of the training data create



(c)

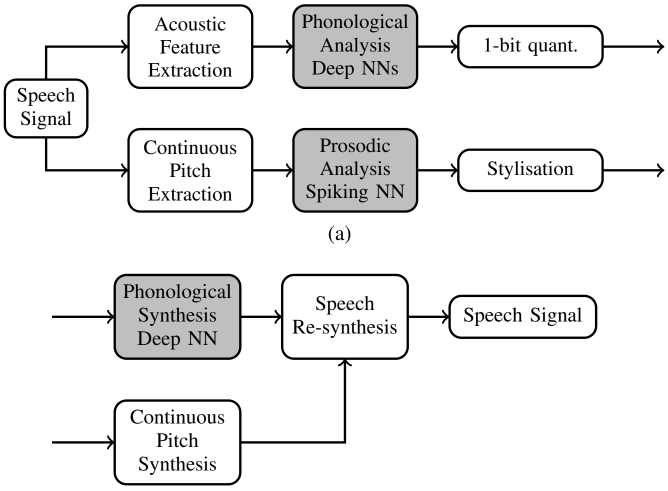
Fig. 1. An illustration of transmission scheme of a demonstration sample shown on 1a, composed of the segmental information shown on 1b: the binary phonological code, and prosodic information shown on 1c: the syllable-based 3bit-quantized second order DLOP parameters.

the segmental codebook. The number of unique binary patterns (the size of the segmental codebook) is a small fraction of the whole permissible patterns (for example, for the eSPE phonological system, it is about 0.5%). The binary patterns are often repeated frame by frame. The segmental code thus consists of an index of the codebook, along with the duration of the code. The output of the spiking NN syllable analysis is stylized using 3-bit quantization of second order DLOP parameters. All stylized F0 mean and F0 slope values create the prosodic codebooks, and the prosodic code consists of the two indexes of the F0 mean and slope codebooks, along with the duration of the transmitted syllable.

The prosodic code is extracted by spiking NN and DLOP parametrization independently of the segmental code. That means that the prosodic code is encoded asynchronously to the segmental one (both codes might have different start, end, and duration). Both segmental and prosodic codes are transmitted in parallel.

To confirm the feasibility of the proposed NN speech coder, we created an experimental framework. The BSD-licensed open-source prototype[1], including pre-trained NNs, is based on:

- phonological vocoding performing speech analysis and synthesis using phonological posteriors [56],
- in-house Idiap LPC vocoding, and



(a)

(b)

Fig. 2. The composition of the functional components of the NN speech coder. Three different NNs, shown in grey, are used: (i) phonological analysis DNNs [46] with 1-bit quantization encoding the binary segmental code, (ii) a spiking NN with stylisation encoding continuous F0, and (iii) a synthesis DNN [46] that decodes the segmental code to the speech parameters.

- syllable onset detection and DLOP-based parametrization.

B. Data

The DNN encoder is trained on the *si_tr_s_284* set of the Wall Street Journal (WSJ0 and WSJ1) continuous speech recognition corpora [57], and the SNN is trained on a subset of the TIMIT corpus [58] (the 10 sentences for speakers indexed 1–100).

The DNN decoder is trained on the Nancy database provided by the Blizzard Challenge[2]. The speaker is known as “Nancy”; she is a US English native female speaker. The database consisted of 16.6 hours of high quality recordings of natural expressive human speech recorded in an anechoic chamber at a 96 kHz sampling rate during 2007 and 2008. The database comprised of 12 095 utterances, and the following split was used:

- the training set, utterances from 1 to 10 000,
- the cross-validation set, utterances from 10 001 to 11 000,
- the test set, the remaining 1095 utterances.

The text was processed by a conventional and freely available TTS front-end [59], and the resulting phonetic labels were used for training of the synthesis DNN.

The same data was also used to train a baseline HMM-based VLBR speech coding system.

C. Training

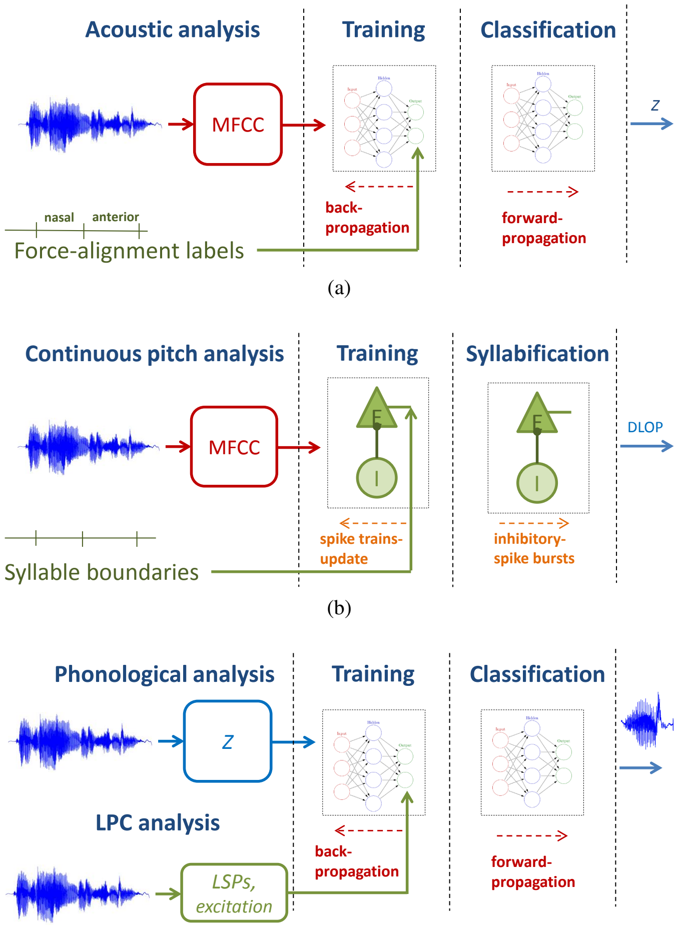
1) Baseline HMM-based system: We trained the baseline HMM-based system as described in [7]. For HMM analysis models, we trained three-state, cross-word triphone models with the HTS variant [60] of the HTK toolkit on the WSJ training set. We tied triphone models with decision tree state

[1] https://github.com/idiap/phonvoc/vlbr

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

clustering based on the minimum description length (MDL) criterion. The MDL criterion allows an unsupervised determination of the number of states. In this study, we obtained 12,685 states each modelled with a GMM consisting of 16 Gaussians. We used mel-frequency cepstral coefficients as acoustic features. The phoneme set comprising 40 phonemes (including “sil”, representing silence) was defined by the CMU pronunciation dictionary.

For building the HMM synthesis models, the implementation of training from the EMIME project [61] was used. Five-state, left-to-right, no-skip HSMMs were used. The speech parameters that were used for training the HSMMs were 39-order cepstral coefficients, log-F0 and 21-band aperiodicities, along with their delta and delta-delta features, framed by 25 ms windows, extracted every 5 ms. Cepstral instead of mel-cepstral features were used, as re-synthesis without mel-warping was almost two times faster.



(a)

(b)

(c)

Fig. 3. Training and inference stages for an analysis DNN shown on 3a, a spiking NN shown on 3b, and a synthesis DNN shown on 3c. 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.

2) Phonological analysis DNNs: Figure 3 shows the training and inference stages of the three different NNs. For phonological analysis DNN training, we first aligned the WSJ training data using HMM acoustic models trained for the baseline system. Then, considering the three different phonological systems, GP, SPE and eSPE, the three different banks of DNNs were trained on the 90% subset of the training set, and the remaining 10% subset was used for cross-validation. For all systems, the labels of phonemes were mapped to the respective phonological classes. In total, K DNNs (12 for GP, 15 for SPE and 21 for eSPE) were trained as phonological analyzers using the short segment (frame) alignment, with two output labels indicating whether the k-th phonological class exists for the aligned phoneme or not. The architecture of the DNNs was $351 \times 1024 \times 1024 \times 1024 \times 2$ neurons, determined empirically. The input vectors were 39-order MFCC features with a temporal context of 9 successive frames.

The training was initialized using deep belief network pre-training done by the single-step contrastive divergence (CD-1) procedure of [62]. The DNNs with the softmax output function were then trained using a mini-batch based stochastic gradient descent algorithm with the cross-entropy cost function of the Kaldi toolkit [63]. Tables I, II and III list the phonological classes and detection accuracy for GP, SPE and eSPE respectively. The DNN outputs for individual phonological classes determine the phonological posterior probabilities.

TABLE I
Phonological classes and classification accuracy of the GP analysis at frame level.

Phonolog. features	Accuracy (%)		Phonolog. features	Accuracy (%)	
	train	cv		train	cv
A	95.2	93.6	i	97.8	96.9
a	98.6	98.0	N	98.9	98.4
E	95.4	93.7	S	97.0	95.8
H	97.0	95.9	u	98.7	98.0
h	97.4	96.4	U	96.7	95.4
I	96.7	95.7	silence	99.4	99.1

TABLE II
Phonological classes and classification accuracy of the SPE analysis at frame level.

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

3) Prosodic analysis SNN: The prosodic analysis SNN is based on an interconnected network composed of 10 excitatory and 10 inhibitory leaky integrate-and-fire neurons. Its principles are based on findings on the role of slow neural oscillations in the auditory cortex for natural speech parsing [64].

The 13-dim cepstral input vectors are first transformed to unidimensional time series. We use weighting to allow for some channels/frequencies to have higher importance, e.g. frequencies around formants likely to provide more information about syllable boundaries because of the vocalisation process. Syllable boundaries are characterised by local minima of the

TABLE III
Phonological classes and classification accuracy of the eSPE analysis at frame level.

Phonolog. features	Accuracy (%)		Phonolog. features	Accuracy (%)	
	train	cv		train	cv
anterior	96.7	95.5	low	98.5	97.7
approximant	98.3	97.4	mid	96.3	95.0
back	96.5	95.1	nasal	98.9	98.4
continuant	97.4	96.4	retroflex	99.1	98.7
coronal	96.9	95.7	round	97.3	96.2
dental	99.6	99.4	stop	98.0	97.2
fricative	98.4	97.8	tense	94.7	92.5
glottal	99.9	99.8	velar	99.5	99.1
high	96.7	95.4	voiced	96.8	95.8
labial	98.9	98.1	vowel	96.5	95.3

weighted signal; these can be generalised to a convolution of the temporal kernel and the weighted signal [54].

Training of the parameters of the spiking NN is based on minimising the syllabic distance of actual syllable boundaries and those produced by the convolution, over the 1000 sentences of the training set. The syllabification program tsylb2 [65] was used to convert phonetician-labelled phonemes and phoneme boundaries into syllables and syllable boundaries.

The output of the spiking NN is stylized using 3-bit quantization of 2^{nd} order DLOP parameters. To create the codebooks, the logarithm of the continuous pitch of all the syllables of the Nancy training data was parametrized with DLOP parameters. The values were linearly spaced between the $\mu - 3\sigma$ and $\mu + 3\sigma$ boundaries, where μ is the mean and σ is the standard deviation of all the measurements of the DLOP parameters. One codebook was thus created for the 1^{st} order DLOP parameter, and one for the 2^{nd} one.

4) Phonological synthesis DNNs: The speech signals from the training and cross-validation sets of the Nancy database, down-sampled to 16 kHz, framed by $25\,\mathrm{ms}$ windows with the three different frame shifts: 10, 16 and $20\,\mathrm{ms}$, were used for extracting both DNN input and output features. The input features, phonological posteriors $\vec{z}_n$, were generated by the phonological analysers. The temporal context of 11 successive frames resulted in input features of $12 \times 11 \times 1 = 132$, $15 \times 11 \times 1 = 165$ and $21 \times 11 \times 1 = 231$ dimensions, for the GP, SPE and eSPE schemes, respectively. The output features, the LPC speech parameters, were extracted by SSP: $\vec{p}_n$ - Line Spectral Pairs (LSPs) of 24th order plus gain, $\log(\vec{r}_n)$ - a Harmonic-To-Noise (HNR) ratio, and $\vec{t}_n$, $\log(\vec{m}_n)$ - two glottal model parameters [66], angle t and magnitude $\log(m)$ of a glottal pole, respectively. Thus, we used static speech parametrization of 28th order along with its dynamic features, altogether of 84th order.

Cepstral mean normalisation of the output features was applied before DNN training. Altogether, 6 DNNs were trained for 3 different phonological schemes and 3 time shifts. The DNNs were initialised using $(K * 11) \times 1024 \times 1024 \times 1024 \times 1024 \times 84$ pre-training and DNNs with a linear output function were then trained by Kaldi with mean square error cost function.

IV. Evaluation

In this section, we present an evaluation of the VLBR phonetic and phonological NN speech coder, and a comparison to the baseline HMM-based system. We encoded and decoded 1095 utterances of the Nancy database test set. In following sections, we present results focusing on:

1) **Phonetic NN speech coder**, comparing the phonetic and phonological coding in Section IV-A.
2) **Phonological NN speech coder**, quantifying an impact of different phonological schemes and frame shifts on speech quality and transmission rates in Section IV-B.
3) **HMM vesus NN speech coders**, comparing the VLBR HMM-based and NN speech coding in Section IV-C.
4) **Intelligibility of NN speech coder** in Section IV-D.

A. Phonetic NN speech coder

Segmental information of the NN speech coder may consist of either phonetic or phonological posteriors (cf. Section II-B), transmitted frame by frame. Therefore, we started evaluation by comparison of the speech quality of the phonetic and phonological speech coding. To measure the impact of the phonetic and phonological posteriors only, the F0 encoding was by-passed with the original F0 signals. To achieve VLBR, binary posteriors have to be used, hence we were interested which of binary phonetic or phonological posteriors perform better.

We normalised the phonetic and phonological posteriors to the binary values (the probabilities above 0.5 are normalized to 1 and the probabilities less than 0.5 are forced to zero) and used Mel Cepstral Distortion (MCD) [67] between original and encoded speech samples as an objective metric to compare overall speech quality. Lower MCD values indicate higher speech quality of the encoded speech samples. The segmental information in this experiment consists of $10\,\mathrm{ms}$ framed vectors of either phone posterior probabilities or eSPE phonological-class posterior probabilities.

TABLE IV
Objective quality evaluation of continuous and binary parametric phonetic and phonological (eSPE) vocoding, using Mel Cepstral Distortion (MCD). MCD is a measure of error-magnitude, whereby smaller is better.

Type / MCD [dB]	Re-synthesis	Continuous	Binary
LPC re-synthesis	4.75	–	–
Phonetic vocoding	–	6.11	6.70
Phonological vocoding	–	6.20	6.46

Table IV reports the results. All results are statistically significant ($p < 0.01$ of a t-test). The first row shows speech distortion caused by the LPC vocoding. The second and third rows report the distortions of phonetic and phonological vocoding done on the top of the LPC vocoding. We can conclude that the majority of the distortion comes from the parametric vocoder; the distortion of phonetic/phonological vocoding with continuous features is about 1.4 dB. While the performance of continuous phonetic and phonological posteriors is similar, feature normalization has higher negative impact on the binary phonetic posteriors. The reason why binary phonological posteriors outperform the phonetic ones could be in the parallel nature of phonological vocoding. Also, maximum a-posteriori classification of phonological posteriors is far more accurate than phonetic posteriors. Therefore, we conclude that phonological posteriors are more suitable for the NN coder, and we selected the binary phonological features in further evaluation.

B. *Phonological NN speech coder*

We used the binary phonological posterior features, and created a codebook from the unique patterns for each phonological scheme. Linear 3-bit quantized syllable-based F0 parametrization is used in this experiment. The 3-bit quantization degrades speech quality by only about 0.1 dB.

Lower bit-rates can be achieved by using a lower dimensional phonological scheme. Figure 4 shows a linear dependence of the codebook size on the number of phonological classes. Increasing the frame shift slightly decreases the number of unique patterns. We speculate that the number of unique binary phonological posteriors is related to the number of clustered senones (tied context-dependent HMM states) used in ASR and TTS acoustic modelling. Then we could interpret a vector of phonological posteriors as a senone.

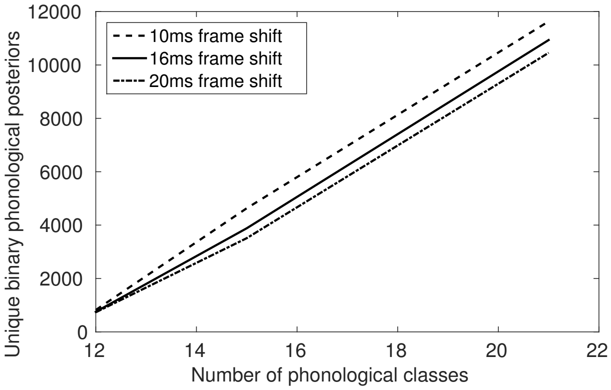


Fig. 4. Linear dependence of unique binary phonological posteriors on the number of phonological classes.

Recall that we consider three different phonological systems in this work. Each system defines a different set of features, with the dimensions 12, 15 and 21, for the GP, SPE and eSPE phonological systems respectively. We discuss their meaning and exact definition in [9].

Figure 5 shows the objective evaluation of the NN speech coding. The transmission rate depends on two variables: the phonological scheme that results in smaller or larger codebooks, and the frame shift. The eSPE codebook consists of 11 632 unique binary phonological patterns; 14 bits are thus required to transmit the segmental code. Fewer bits are required for the SPE and GP codebook, 12 and 10 bits respectively. We can see from the figure that the difference in quality degradation lies in the range of about 0.2 dB, therefore we identified the GP system as the most suitable for the VLBR system. The frame shift has a bigger impact on the bit rate; for the GP system, when increasing the frame shift from 10 ms to 16 ms, the degradation increases by 0.3 dB. We consider the frame shift of 16 ms as a good balance between the transmission rate and the speech quality.

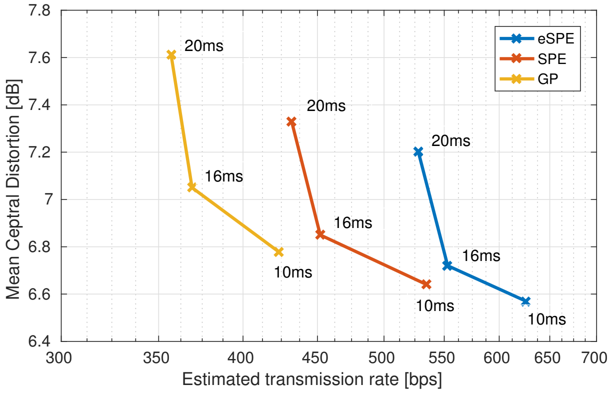


Fig. 5. Estimated transmission rates of the NN speech coding. Data points are estimated for different phonological systems and various frame shifts.

The overall quality of the NN speech coding was evaluated subjectively using the Degradation Category Rating (DCR) procedure [68] quantifying the Degradation Mean Opinion Score (DMOS). The aim was to estimate speech encoding quality variations based on the different frame-shift. The test consisted of 20 randomly selected utterances from the Nancy test data, at least 2 seconds long. 37 listeners were asked to rate the degradation of re-synthesised signals, compared with reference signals, based on their overall perception. Figure 6 shows that NN speech coding, operating at 360–370 bit/s, achieves a MOS of approximately 2.3. The figure also shows higher uncertainty of listeners (higher standard deviation of MOS) when increasing the frame shift of the DNN coder. Although similar DMOS is achieved with both the 16 ms and 20 ms frame shifts, speech coding with the 20 ms frame shift has much higher standard deviation of subjective testing. Thus, we selected the GP-based 16 ms frame shift system as an optimal VLBR NN speech coder for a qualitative and intelligibility comparison with the HMM coder.

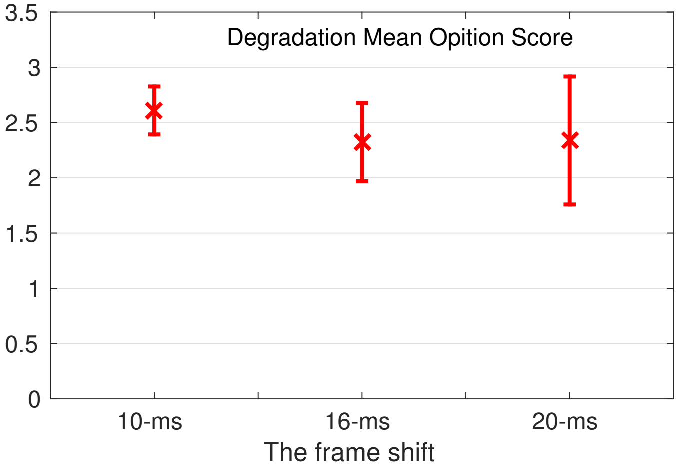


Fig. 6. Subjective evaluation of DNN-based speech coding using the GP phonological scheme with various frame shifts. Degradation categories are: 3 – slightly annoying, 2 – annoying, and 1 – very annoying.

C. *HMM versus NN coding*

The HMM coder creates bursts of segmental errors when phoneme ASR fails; this is certainly disruptive to the listeners. It also uses voiced/unvoiced detection for parametrization of the F0 signal only in voiced regions. This prosodic encoding plugged-in to the VLBR system creates additional speech discontinuities and unnatural speech artefacts. On the other hand, the NN speech coder transmits information per frame, and parameterizes continuous F0, that altogether significantly reduces discontinuities during speech reconstruction at the receiver. Therefore, we continued in perceptual comparison of the HMM and NN coding.

For this qualitative comparison, we employed a 5-point scale ABX subjective evaluation listening test [69]. In this test, listeners were presented with pairs of samples produced by two systems (A and B) and for each pair they indicated their preference or strong preference for A, B, or *both samples sound the same* (X). The material for the test consisted of 20 pairs of sentences such that one member of the pair was generated using the GP-based 16 ms frame shift NN speech coder (system A), and the other member was generated using the HMM-based speech coder (system B). Random utterances from the test set of the Nancy database were used to generate the encoded speech. We chose this GP system as a compromise between the speech quality and the low bit rate. The subjects for the ABX test were 32 native English listeners, roughly equally pooled from experts in speech processing on the one hand, and completely naive subjects on the other hand. The subjects were presented with pairs of sentences in a random order with no indication of which system they represented. They were asked to listen to these pairs of sentences (as many times as they wanted), and choose between them in terms of their overall quality. Additionally, the option X, i.e. *both samples sound the same*, was available if they had no preference for either of them.

As can be seen in Figure 7, the NN speech coder significantly outperforms the HMM-based one. The strong preference and preference choices of the NN coder achieve 30.62% and 54.22% (sum up to 84.84%) over 0.63% and 4.53% (sum up to 5.16%) respectively for the HMM-based one. In addition the "no preference" choice achieved 10%.

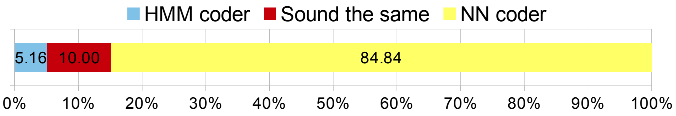


Fig. 7. ABX subjective evaluation test of HMM and DNN/SNN coders.

D. *Intelligibility of NN coder*

Finally, we were interested in intelligibility degradation of NN coding. As a baseline, we selected the Open Source/Free Software speex codec[3]. Speex is based on CELP and is designed to compress voice at bitrates ranging from 2.15 to 44 kbit/s. It also allows wideband (16 kHz sampling rate) speech compression, the rate with which we designed the NN coding.

We evaluated HMM, NN coding and the Speex coding using the short-time objective intelligibility (STOI) measure [70]. STOI analyzes the short-time magnitudes of signals within one-third octave bands, and computes the correlations between reference and test signal band envelopes over segments of 384 ms. The individual correlation values are then averaged over bands and segments to yield an objective score in the range [0, 1] that is expected to have a monotonic relation with intelligibility. This approach was found to provide good predictions of subjective intelligibility scores for speech coding as used in cellular networks [71]. We used the publicly available MATLAB implementation [72].

TABLE V

The STOI intelligibility evaluation of HMM coding operating at 213 bit/s, *NN coding operating at 357–626* bit/s, *and the Speex coding, operating at 2150 bps.*

Codec type	HMM based	NN / 10 ms	NN / 16 ms	NN / 20 ms	Speex codec
GP	0.38	0.70	0.68	0.64	0.78
SPE		0.71	0.70	0.66	
eSPE		0.71	0.71	0.67	

Table V shows the intelligibility comparison of HMM, NN and the Speex coding. We selected the minimal transmission rate offered by the Speex coder to compare different coding techniques at the most similar bit rates. We can conclude that NN coding significantly outperforms the HMM coding in terms of intelligibility[4] and, compared to LPC coding, it degrades intelligibility by only 10% (for example for the GP-based 16 ms frame shift NN speech coder); however, it operates at a bit rate approximately 6 times lower.

V. Bit Allocation and Complexity

The test set contained about 36 000 syllables (syll) in 5240 s of speech including silences. On average, there were 6.8 syll/s. The transmission code included the index of the binary pattern along with its duration, the segmental code, and two indexes of quantized mean and slope of syllable-based F0 codebooks, along with the syllable duration and the supra-segmental code. Both segmental and suprasegmental information are transmitted asynchronously, i.e., the segmental blocks and syllables have different start, end and duration.

As an example of the bit allocation, Table VI shows the details for the GP system used in the ABX test. Closer analysis of repeated patterns of the segmental code reveals that the number of blocks is less than 46% of the total number of frames and 2 bits is sufficient to transmit the number of repeated codes. That amounts to $0.46 \times 56 \times 10 = 257$ bit/s transmission rate for the GP codes, and $0.46 \times 56 \times 2 = 52$ bit/s transmission rate for the segmental code duration. Because around 10% of the transmitted speech is detected as silence (valid for our training

[3] http://www.speex.org/

[4] The lower intelligibility of HMM speech coding could be caused by less precise phoneme boundary detection performed by the ASR system, which resulted in segmental misalignment of the encoded and reference speech samples and that impacted the objective intelligibility measure.

TABLE VI
GP-BASED NEURAL NETWORK VOCODER BIT ALLOCATION.

Parameter	bit/unit	Unit	bit/s
GP code	10	Code block	257
Duration	2	Code block	52
F0 mean	3	Syllable	18
F0 slope	3	Syllable	18
Duration	4	Syllable	24
Total			**369**

data), we obtain the effective speech frame-rate for the $16\,\mathrm{ms}$ frame shift as $62.5 \times 0.1 = 56$ bit/s. The transmission rate of supra-segmental code is constant for all tested NN-based speech coding systems, as the number of syllables is always constant. It consists of an index of the F0 mean codebook ($3 \times 6 = 18$ bit/s), and index of the F0 slope codebook ($3 \times 6 = 18$ bit/s), and the syllable duration (4 bit/s). The effective syllable rate (leading and trailing syllables removed) of our data was $6\,\mathrm{syll/s}$. The duration of syllables is encoded by $4\,\mathrm{bit}$ (covering duration up to $256\,\mathrm{ms}$) that results in encoding supra-segmental information at a constant 60 bit/s. Altogether, the estimated bit rate for the NN-based speech coding using the GP phonological scheme and the $16\,\mathrm{ms}$ frame shift is 369 bit/s.

The average syllable duration, including leading, trailing and short pause silences, was $150\,\mathrm{ms}$. As both speech encoding and decoding processing (forward passes of the NNs) were faster than real-time, we consider the average syllable duration as an algorithmic latency of $150\,\mathrm{ms}$ of the proposed coder.

According to the G.114, the users are “very satisfied” as long as latency does not exceed $200\,\mathrm{ms}$ [73]. G.114 refers to total latency of the complete system including algorithmic- and transmission-latency. For example the 3GPP Enhanced Voiced Services codec constrains algorithmic latency to $32\,\mathrm{ms}$, such that the expected network-latency of around $150\,\mathrm{ms}$ can be contained within the total $200\,\mathrm{ms}$ limit. We estimate that the total latency of our method leaving $130\,\mathrm{ms}$ for network-latency should not exceed $280\,\mathrm{ms}$, for which G.114 rates users as “satisfied”.

A. Complexity of NNs

Since most coders are assumed to be used in portable devices, we evaluated also the suitability of NNs for small footprint systems.

Running the NN speech coder based on the GP system requires 12 analysis DNNs, 1 synthesis DNN and 1 SNN on a device. Since the SNN is a small footprint system that consists of just 10 excitatory and 10 inhibitory leaky integrate-and-fire neurons, we focus rather on DNNs. Each analysis DNN was trained with 2.46 million parameters that would require about $10\,\mathrm{MB}$ memory (parameters stored in single-precision floating-point format). The synthesis DNN was trained with 3.31 million parameters that would require $13.2\,\mathrm{MB}$ memory. All DNNs thus require $132\,\mathrm{MB}$.

The computational complexity of the proposed end-to-end NN speech coder depends mostly on the size of NNs. The speech encoding and decoding is realised as a single NN forward pass, and thus its computational complexity is about N_w; that is, the number of weights of the NNs.

Recent research on complexity reduction of NNs (for example [74]–[77]) showed a possible 80% memory reduction. We then estimate that the reduced memory could fit the runtime memory of Digital Signal Processors (DSPs) available in typical smart-phone hardware [78]. Effective methods for small footprint development rely on exploiting the structured representation of weight matrices [76]. The low-rank factorization and Toeplitz matrices are found to enable significant memory reduction with minor or no reduction in performance [75], [76]. Furthermore, the knowledge in a large DNN can be distilled in a smaller size DNN if the output of the large DNN is used as the labels (instead of the hard labels) for training the small size DNN [74]. We plan to explore the idea of soft target training in the context of DNN based speech coding in our future studies.

VI. CONCLUSIONS

VLBR speech coding based on a recognition/synthesis paradigm is normally either corpus-based (using unit-selection approach), or HMM-based (using HMM-based ASR or TTS). We have designed and presented a NN-based speech coding composed of deep NNs and a spiking NN; the solution represents an end-to-end neural network based VLBR speech coding.

We have compared phonetic and phonological NN coding; given the binary nature of the phonological posteriors, they outperform binary phonetic posteriors. Further, we have compared three different phonological systems, and we conclude that an optimal NN speech coder can be designed by using the phonological posteriors defined by the Government Phonology classes. By selecting a frame shift of $16\,\mathrm{ms}$, the NN coder operates at 369 bit/s with a latency of $150\,\mathrm{ms}$.

Listener preference evaluation of HMM and NN-based speech coders showed that a NN speech coder with continuous F0 modelling is significantly preferred by (85% of) the listeners. We have concluded that NN coding significantly outperforms the HMM coding in terms of intelligibility. As we have used an open-source experimental framework with a rather standard LPC vocoder, there is potential for higher speech quality from NN speech coding in future by improved parametric vocoding. Table IV shows that more than 77% of all degradation comes from the parametric vocoding.

The design of the proposed coder is simplified to just the three kinds of neural networks. Our future work will be focused on investigations of computational complexity reduction of the NN speech coding, and further speech quality improvements targetting quality of LPC coding at 2 kbit/s. We believe that this coding approach will then become more feasible for a range of computation platforms that may be used in military and tactical communication systems.

ACKNOWLEDGMENT

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS), and under SP2: the SCOPES Project on Speech Prosody. The wor was also partly supported under the RECOD project by armasuisse, the Procurement and Technology Center of the Swiss Federal Department of Defence, Civil Protection and Sport.

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

References

[1] S. Dimolitsas, C. Ravishankar, and G. Schroder, “Current objectives in 4-kb/s wireline-quality speech coding standardization,” *IEEE Signal Processing Letters*, vol. 1, no. 11, pp. 157–159, Nov. 1994.

[2] K.-S. Lee and R. Cox, “A very low bit rate speech coder based on a recognition/synthesis paradigm,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 9, no. 5, pp. 482–491, Jul 2001.

[3] G. V. Baudoin and F. El Chami, “Corpus based very low bit rate speech coding,” in *Proc. of ICASSP*, vol. 1. IEEE, Apr. 2003, pp. I–792–I–795 vol.1.

[4] J. Picone and G. R. Doddington, “A phonetic vocoder,” in *Proc. of ICASSP*. IEEE, May 1989, pp. 580–583 vol.1.

[5] K. Tokuda, T. Masuko, J. Hiroi, T. Kobayashi, and T. Kitamura, “A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques,” in *Proc. of ICASSP*, vol. 2. IEEE, May 1998, pp. 609–612 vol.2.

[6] A. McCree, K. Brady, and T. F. Quatieri, “Multisensor very lowbit rate speech coding using segment quantization,” in *Proc. of ICASSP*. IEEE, Mar. 2008, pp. 3997–4000.

[7] 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.

[8] G. Greenberg and S. Chang, “Linguistic dissection of switchboard-corpus automatic speech recognition systems,” in *Proc. of ITRW on Automatic Speech Recognition: Challenges for the new Millenium*, Paris, France, 2000, pp. 195–202.

[9] M. Cernak, S. Benus, and A. Lazaridis, “Speech vocoding for laboratory phonology,” 2016. [Online]. Available: http://arxiv.org/abs/1601.05991

[10] J. Harris, *English Sound Structure*, 1st ed. Wiley-Blackwell, Dec. 1994.

[11] J. Harris and G. Lindsey, *The elements of phonological representation*. Harlow, Essex: Longman, 1995, pp. 34–79.

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

[13] D. Yu, S. Siniscalchi, L. Deng, and C.-H. Lee, “Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition,” in *Proc. of ICASSP*. IEEE SPS, March 2012.

[14] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, “Experiments on Cross-Language Attribute Detection and Phone Recognition With Minimal Target-Specific Training Data,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 875–887, Mar. 2012.

[15] 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.

[16] Y. Bengio, “Learning deep architectures for AI,” *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[17] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994, iSBN 0-7923-9396-1.

[18] A. Mohamed, G. E. Dahl, and G. E. Hinton, “Deep belief networks for phone recognition,” in *NIPS’22 workshop on deep learning for speech recognition*, 2009.

[19] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[20] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing (receiving 2013 IEEE SPS Best Paper Award)*, vol. 20, no. 1, pp. 30–42, January 2012.

[21] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using Deep Neural Networks,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7962–7966.

[22] S. Esmeir, S. Markovitch, and C. Sammut, “Anytime learning of decision trees,” *Journal of Machine Learning Research*, vol. 8, pp. 891 – 933, 2007.

[23] K. Yu, H. Zen, F. Mairesse, and S. Young, “Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis,” *Speech Communication*, vol. 53, no. 6, pp. 914–923, Jul. 2011.

[24] Y. Qian, Y. Fan, W. Hu, and F. Soong, “On the training aspects of deep neural network (DNN) for parametric tts synthesis,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3829–3833.

[25] A. Lazaridis, B. Potard, and P. N. Garner, “DNN-based speech synthesis: Importance of input features and training data,” in *International Conference on Speech and Computer, SPECOM 2015*, ser. Lecture Notes in Computer Science, N. F. A. Ronzhin, R. Potapova, Ed. Springer Berlin Heidelberg, 2015, pp. 193–200.

[26] B. S. Atal and J. R. Remde, “A new model of LPC excitation for producing natural-sounding speech at low bit rates,” in *Proc. of ICASSP*, vol. 7. IEEE, May 1982, pp. 614–617. [Online]. Available: http://dx.doi.org/10.1109/icassp.1982.1171649

[27] M. Schroeder and B. Atal, “Code-excited linear prediction(CELP): High-quality speech at very low bit rates,” in *Proc. of ICASSP*, vol. 10. IEEE, Apr. 1985, pp. 937–940. [Online]. Available: http://dx.doi.org/10.1109/icassp.1985.1168147

[28] S. Morishima, H. Harashima, and Y. Katayama, “Speech coding based on a multi-layer neural network,” in *Communications, 1990. ICC &#039;90, Including Supercomm Technical Sessions. SUPERCOMM/ICC &#039;90. Conference Record., IEEE International Conference on*. IEEE, Apr. 1990, pp. 429–433 vol.2. [Online]. Available: http://dx.doi.org/10.1109/icc.1990.117117

[29] Y. Zhen, “Prediction in speech coding: the modification of the coding of LPC parameters and nonlinear estimation technique by using ANN,” in *Signal Processing, 1996., 3rd International Conference on*, vol. 1. IEEE, Oct. 1996, pp. 690–693 vol.1. [Online]. Available: http://dx.doi.org/10.1109/icsigp.1996.567357

[30] S. Hunt, “A nonlinear adaptive predictor for speech compression,” in *Neural Networks, 1996., IEEE International Conference on*, vol. 4. IEEE, Jun. 1996, pp. 1998–2002 vol.4. [Online]. Available: http://dx.doi.org/10.1109/icnn.1996.549208

[31] C. Chavy, B. Gas, and J. L. Zarader, “Discriminative coding with predictive neural networks,” in *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, vol. 1. IET, 1999, pp. 216–220 vol.1. [Online]. Available: http://dx.doi.org/10.1049/cp:19991111

[32] M. Faúndez-Zanuy, *Engineering Applications of Bio-Inspired Artificial Neural Networks: International Work-Conference on Artificial and Natural Neural Networks, IWANN’99 Alicante, Spain, June 2–4, 1999 Proceedings, Volume II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, ch. Adaptive hybrid speech coding with a MLP/LPC structure, pp. 814–823. [Online]. Available: http://dx.doi.org/10.1007/BFb0100549

[33] B.-K. Lee and J.-H. Chang, “Packet Loss Concealment Based on Deep Neural Networks for Digital Speech Transmission,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 378–387, Feb. 2016. [Online]. Available: http://dx.doi.org/10.1109/taslp.2015.2509780

[34] M. Sheikhan, V. T. Vakili, and S. Garoucy, “Complexity Reduction of LD-CELP Speech Coding in Prediction of Gain Using Neural Networks,” *World Applied Sciences Journal*, no. 7, pp. 38–44, 2009.

[35] M. G. Easton and C. C. Goodyear, “A CELP codebook and search technique using a Hopfield net,” in *Proc. of ICASSP*. IEEE, Apr. 1991, pp. 685–688 vol. 1. [Online]. Available: http://dx.doi.org/10.1109/icassp.1991.150432

[36] L. Wu, M. Niranjan, and F. Fallside, “Fully vector-quantized neural network-based code-excited nonlinear predictive speech coding,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 2, no. 4, pp. 482–489, Oct. 1994. [Online]. Available: http://dx.doi.org/10.1109/89.326608

[37] D. W. Griffin and J. S. Lim, “Multiband excitation vocoder,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 36, no. 8, pp. 1223–1235, Aug. 1988. [Online]. Available: http://dx.doi.org/10.1109/29.1651

[38] H. Cui and H. Jiang, “A robust 800 bps MBE coder with VQ and MLP,” in *Communication Technology Proceedings, 1998. ICCT &#039;98.*

1998 International Conference on, vol. vol.2. IEEE, Oct. 1998, pp. 4 pp. vol.2+. [Online]. Available: http://dx.doi.org/10.1109/icct.1998.741011

[39] 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. [Online]. Available: http://dx.doi.org/10.1109/tassp.1982.1163960

[40] 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. [Online]. Available: http://dx.doi.org/10.1109/icassp.1982.1171472

[41] ——, "A segment vocoder at 150 b/s," in *Proc. of ICASSP*, vol. 8. IEEE, Apr. 1983, pp. 61–64. [Online]. Available: http://dx.doi.org/10.1109/icassp.1983.1172241

[42] 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. [Online]. Available: http://dx.doi.org/10.1109/icassp.1983.1172244

[43] 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. [Online]. Available: http://dx.doi.org/10.1109/tassp.1985.1164584

[44] 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. [Online]. Available: http://dx.doi.org/10.1109/29.90372

[45] J. Gibson, "Speech Compression," *Information*, vol. 7, no. 2, pp. 32+, Jun. 2016. [Online]. Available: http://dx.doi.org/10.3390/info7020032

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

[47] R. E. Turner and M. Sahani, "Demodulation as Probabilistic Inference," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2398–2411, Nov. 2011.

[48] V. Leong, M. A. Stone, R. E. Turner, and U. Goswami, "A role for amplitude modulation phase relationships in speech rhythm perception." *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 366–381, Jul. 2014.

[49] P. Lakatos, A. S. Shah, K. H. Knuth, I. Ulbert, G. Karmos, and C. E. Schroeder, "An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex." *Journal of neurophysiology*, vol. 94, no. 3, pp. 1904–1911, Sep. 2005.

[50] M. Cernak, A. Asaei, and H. Bourlard, "On Structured Sparsity of Phonological Posteriors for Linguistic Parsing," 2016. [Online]. Available: http://arxiv.org/abs/1601.05647

[51] K. Yu, B. Thomson, S. Young, and T. Street, "From Discontinuous To Continuous F0 Modelling In HMM-based Speech Synthesis," in *Proc. ISCA SSW7*. Kyoto, Japan: ISCA, 2010, pp. 94–99.

[52] K. Yu and S. Young, "Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, Jul. 2011. [Online]. Available: http://dx.doi.org/10.1109/tasl.2010.2076805

[53] K. U. Ogbureke, J. P. Cabral, and J. Carson-Berndsen, "Using Noisy Speech to Study the Robustness of a Continuous F0 Modelling Method in HMM-based Speech Synthesis," in *Proc. Speech Prosody*, Shanghai, China, 2012, pp. 67–70.

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

[55] M. Cernak, X. Na, and P. N. Garner, "Syllable-Based Pitch Encoding for Low Bit Rate Speech Coding with Recognition/Synthesis Architecture," in *Proc. of Interspeech*, Aug. 2013, pp. 3449–3452.

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

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

[58] L. D. Consortium, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Philadelphia, 1993.

[59] A. Black, P. Taylor, and R. Caley, "The Festival Speech Synthesis System," Human Communication Research Centre, University of Edinburgh, Technical Report, 1997.

[60] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based Speech Synthesis System Version 2.0," in *Proc. of ISCA SSW6*, 2007, pp. 131–136.

[61] M. Wester, J. Dines, M. Gibson, H. Liang, Y.-J. Wu, L. Saheer, S. King, K. Oura, P. N. Garner, W. Byrne, Y. Guan, T. Hirsimäki, R. Karhila, M. Kurimo, M. Shannon, S. Shiota, J. Tian, K. Tokuda, and J. Yamagishi, "Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project," in *SSW7*, 2010, pp. 192–197.

[62] G. E. Hinton, S. Osindero, and Y. W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

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

[64] 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. [Online]. Available: http://dx.doi.org/10.7554/elife.06213

[65] W. M. Fisher, "tsylb2," 1996. [Online]. Available: http://www.nist.gov/speech/tools

[66] P. N. Garner, M. Cernak, and B. Potard, "A simple continuous excitation model for parametric vocoding," in *Proc. of ICASSP*. IEEE, Apr. 2015, in review.

[67] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. of ICASSP*, vol. 1. IEEE, May 1993, pp. 125–128 vol.1.

[68] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," (Geneva, Switzerland) 1996.

[69] V. Grancharov and W. . B. Kleijn, "Speech Quality Assessment," in *Springer Handbook of Speech Processing*, J. Benesty, Sondhi, and Y. Huang, Eds. Springer Berlin Heidelberg, 2008, pp. 83–100.

[70] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time&#x2013;Frequency Weighted Noisy Speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011. [Online]. Available: http://dx.doi.org/10.1109/tasl.2011.2114881

[71] S. Jørgensen, J. Cubick, and T. Dau, "Speech Intelligibility Evaluation for Mobile Phones," *Acta Acustica united with Acustica*, pp. 1016–1025. [Online]. Available: http://dx.doi.org/10.3813/aaa.918896

[72] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. of ICASSP*. IEEE, Mar. 2010, pp. 4214–4217. [Online]. Available: http://dx.doi.org/10.1109/icassp.2010.5495701

[73] ITU-T Rec. G.114, "One-way transmission time ," (Geneva, Switzerland) 2003.

[74] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," Mar. 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

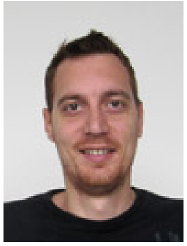
[75] P. Nakkiran, R. Alvarez, R. Prabhavalkar, and C. Parada, "Compressing Deep Neural Networks using a Rank-Constrained Topology," in *Proc. of Interspeech*, 2015, pp. 1473–1477.

[76] V. Sindhwani, T. N. Sainath, and S. Kumar, "Structured Transforms for Small-Footprint Deep Learning," Oct. 2015. [Online]. Available: http://arxiv.org/abs/1510.01722

[77] L. Lu and S. Renals, "Small-footprint deep neural networks with highway connections for speech recognition," *CoRR*, vol. abs/1512.04280, 2015. [Online]. Available: http://arxiv.org/abs/1512.04280

[78] N. D. Lane, P. Georgiev, and L. Qendro, "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*, 2015, pp. 283–294. [Online]. Available: http://doi.acm.org/10.1145/2750858.2804262

Milos Cernak holds a Ph.D. in Telecommunications from Slovak University of Technology in Bratislava. Since 2011 he is a senior engineer and scientific collaborator at Idiap Research Institute, Martigny, Switzerland, involved in various Swiss R&D academic and industrial projects with a core focus on speech recognition and synthesis, low bit rate speech coding, and prosody parametrization. From 2008, he was a member of IBM Research in Prague, working on embedded ViaVoice speech recognition and IBM Reading Companion. After graduating in 2005, he was a post-doc researcher at Institute EURECOM in France, and a principal researcher at Slovak Academy of Sciences. During his studies in 2001, he was also a visiting scientist at Iowa State University's Virtual Reality Application Centre, Ames, USA. He is interested in signal processing, phonology, and neural basis of speech production and perception. He was a plenary lecturer for the 2016 Frederick Jelinek Memorial Summer Workshop.



Alexandros Lazaridis was born in Thessaloniki, Greece, in 1981. He graduated in 2005 (Diploma) from the Department of Electrical and Computer Engineering of the Aristotle University of Thessaloniki, Greece. He received his PhD degree in Feb. 2011 from the Department of Electrical and Computer Engineering of the University of Patras, Greece. Since Nov. 2012 he is a post-doctoral researcher at Idiap Research Institute, Martigny, Switzerland. He is author and co-author in more than 25 publications in scientific journals and international conferences. His research interests include speech and audio signal processing, speech synthesis, speech prosody and spoken language/dialect/accent identification.

Afsaneh Asaei received the B.S. degree from Amirkabir University of Technology and the M.S. (honors) degree from Sharif University of Technology, in Electrical and Computer engineering, respectively. She held a research engineer position at Iran Telecommunication Research Center (ITRC) from 2002 to 2008. She then joined Idiap Research Institute in Martigny, Switzerland as a doctoral research assistant and Marie Curie fellow on speech communication with adaptive learning training network. She received the Ph.D. degree in 2013 from École Polytechnique Fédérale de Lausanne. Her thesis focused on structured sparsity for multiparty reverberant speech processing, and its key idea was awarded the IEEE Spoken Language Processing Grant. Currently, she is a postdoctoral researcher at Idiap Research Institute. She has served as a guest editor of Speech Communication special issue on Advances in sparse modeling and low-rank modeling for speech processing and co-organized special issues on this topic at HSCMA'2014 and LVA/ICA'2015. Her research interests lie in the areas of signal processing, machine learning, statistics, acoustics, auditory scene analysis and cognition, and sparse signal recovery and acquisition.

Philip N. Garner received the degree of M.Eng. in Electronic Engineering from the University of Southampton, U.K., in 1991, and the degree of Ph.D. (by publication) from the University of East Anglia, U.K., in 2012. He first joined the Royal Signals and Radar Establishment in Malvern, Worcestershire working on pattern recognition and later speech processing. In 1998 he moved to Canon Research Centre Europe in Guildford, Surrey, where he designed speech recognition metadata for retrieval. In 2001, he was seconded (and subsequently transfered) to the speech group at Canon Inc. in Tokyo, Japan, to work on multilingual aspects of speech recognition and noise robustness. As of April 2007, he is a senior research scientist at Idiap Research Institute, Martigny, Switzerland, where he continues to work in R&D of speech recognition, synthesis and signal processing. He is a senior member of the IEEE, and has published internationally in conference proceedings, patent, journal and book form as well as serving as coordinating editor of ISO/IEC 15938-4 (MPEG-7 Audio).