SPEECH VOCODING FOR LABORATORY PHONOLOGY

Milos Cernak Alexandros Lazaridis[a]

Idiap-RR-07-2015

MAY 2015

[a] Idiap Research Institute

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

Speech vocoding for laboratory phonology

Milos Cernak, Stefan Benus, Alexandros Lazaridis

May 13, 2015

Abstract

In this paper, we propose a platform based on phonological speech vocoding for examining relations between phonology and speech processing, and in broader terms, between the abstract and physical structures of speech signal. The goal of this paper is to go a step further towards bridging both fields and contributing to the program of Laboratory Phonology. We show two application examples for the laboratory phonology: a comparison of the phonological systems and an experimental phonological parametric text-to-speech (TTS) system. The following three phonological systems are considered in this work: (i) Government Phonology (GP) features, (ii) the Sound Pattern of English (SPE) features, and (iii) the extended SPE (eSPE) features. Comparing GP- and eSPE-based vocoded speech, we conclude that eSPE achieves slightly better results than the other two systems. However, GP – the most compact phonological speech representation – performs comparably to the systems with higher number of phonological classes. The parametric TTS based on phonological speech representation, and trained from an unlabelled audio book in an unsupervised manner, achieves promising results.

We envision that the presented approach paves the way for researchers in both fields to form meaningful hypotheses that are explicitly testable using the concepts developed and exemplified in this paper. Laboratory phonologists might test the applied concepts of their theoretical models and speech processing community may be suggesting tests for the potential of advancing the performance of current state-of-the-art applications using the concepts developed for the theoretical phonological models.

1 Introduction

Speech is a traditional domain exemplifying the dichotomy between continuous and discrete, and in broader terms, between the body and the mind. On the one hand, the articulatory activity and the resulting speech signal are continuously varying. On the other hand, for speech communication to convey meaning, this continuous signal must be, at the same time, perceivable as discretely varying, and thus contrastive. Traditionally, these two aspects have been studied under the labels of phonetics and phonology respectively. Following significant successes of this dichotomous approach, for example in speech synthesis and recognition, in the recent decades a lot of progress have been also seen in understanding and formal modelling of the relationship between these two aspects, e.g. the program of Laboratory Phonology (Pierrehumbert et al., 2000) or the renewed interest in the apparoaches based on Analysis by Synthesis (Hirst, 2011; Bever and Poeppel, 2010). The goal of this paper is to follow these developments by proposing a platform for examining relations between the abstract and physical structures of the speech signal. In this, we aimed at mutual cross-fertilisation between phonology, as a quest for understanding and modelling of cognitive abilities that underlay systematic patterns in our speech, and speech processing, as a quest for natural, robust, and reliable automatic systems for synthesising and recognising speech.

As a first step in this direction we examine vocoding of phonological representations and how this might inform both quests mentioned above. In vocoding — parametric speech coding — sequential processing is traditionally phone-based. The vocoding consists of cascaded phone-based ASR and TTS, as described, e.g., by Picone and Doddington (1989), Tokuda et al. (1998) and Lee and Cox (2001). However, phonological representations of speech have also been shown to be useful for speech processing e.g. in King and Taylor (2000). We expand on this approach by exploring a direct link between phonological features and their engineered acoustic realizations. In other words, we believe that abstract phonological sub-segmental, segmental, and suprasegmental structures may be lawfully related to the physical speech signal through a speech engineering approach, and that this relationship is informative for both phonology and speech processing.

The motivation for this approach is two-fold. First, phonological representations (together with grammar) create a formal phonological model whose overall goal is to capture the core properties of the cognitive system underlying speech production and perception. This unified system of sub-segmental, segmental, and suprasegmental phonological features of speech finds, according to Giraud and Poeppel (2012), independent support in their embodied nature as it correspond to brain-generated cortical oscillation in the 'delta' range (1–3 Hz), 'theta' range (4-7 Hz), and faster 'gamma' range (25–40) Hz that correspond to the stress, syllable and sub-segmental temporal scales, respectively. Hence, speech processing utilizing such a system might lead to a biologically sensible and empirically testable computational model of speech.

Second, phonological representations are inherently multilingual (Siniscalchi et al., 2012). This in turn has an attractive advantage in the context of multilingual speech processing in lessening the reliance on purely phonetic decisions. The language independence of the phonological representations on the one hand and the availability of language specific mapping between these representations and the acoustic signal through speech processing methods on the other hand, offer (we hope) a way to the context-based interpretation of the phonological representation grounded in the phonetic substance yet abstract enough to allow for more streamlined approach to multilingual speech processing.

In this work, we propose to use the phonological vocoding of Cernak et al. (2015) for bridging theoretical phonology and applied computer science. We consider following phonological systems in this work:

- The Government Phonology (GP) features (Harris and Lindsey, 1995) describing sounds by fusing and splitting of 11 primes.
- The Sound Pattern of English (SPE) system with 13 categories established from natural (articulatory) classes (Chomsky and Halle, 1968).
- The extended system of the SPE features (eSPE) (Yu et al., 2012; Siniscalchi et al., 2012) consisting of 21 phonological classes.

Having trained the phonological vocoders for the three phonological models of sound representation, we describe several application examples for this bridging. For example, we test the hypothesis that the best phonological speech representation achieves the best quality vocoded speech, i.e., the phonological features are verified in both directions, recognition and synthesis, simultaneously. Additionally, we compare the segmental properties of the phonological systems, and describe promising results and advantages of experimental phonological parametric TTS synthesis.

The structure of the paper is as follows: the phonological representations used in this work are introduced in Section 2. Section 3 introduces speech vocoding based on phonological speech representation. Section 4 describes the experimental setup. The application examples of the proposed platform (along with the experiments and results) are shown in Section 5. Finally the conclusions follow in Section 6.

2 Phonological systems

Phonology is construed in this work as a formal model that represents cognitive (subconscious) knowledge of native speakers regarding the systematic sound patterns of a language. The two traditional components of such models are i) system primitives, that is, the units of representation for cognitively relevant objects such as sounds or syllables, and ii) a set of permissible operations over these representations that is able to generate the observed patterns. Naturally, these two facets are closely linked and inter-dependent. In this paper we focus on the theory of representation.

2.1 Segmental representations

The minimal unit of meaning contrast, i.e. cognitive relevance, is assumed to be the phoneme. In the tradition of Jakobson and Halle (1956) and Chomsky and Halle (1968), phonemes are assumed to consist of feature bundles. In the former model, 12 basic perceptual-acoustic domains (e.g. acute-grave, or compact-diffuse) define the space for characterising all the phonemes. The model uses polar opposites for these 12 continua, which are necessarily relational, and thus language-specific. Hence, a vowel characterised as, e.g. grave in one language might be phonetically different from the same grave vowel in another language since their grave quality might be at a different points of the acute-grave continuum. The latter system of SPE differed from the former in two fundamental aspects relevant for this paper. First, it took the articulatory production mechanism as the underlying principle of phoneme organisation; hence, in their 13 basic binary features, we talk about the position (or activity) of the active articulators rather than percepts they create. Second, SPE assumed that the flat, unstructured binary feature specifications are language independent and characterise the set of possible phonemes in languages of the world.

The developments of phonological theory after SPE focused on both the theory of representations as well as the operations. The most relevant for this paper are proposals for establishing the non-linear nature of phonological representations. Starting with the representation of lexical tones (Leben, 1973; Goldsmith, 1976), continuing with the featural geometry approach (Sagey, 1986; Clements and Hume, 1995) and receiving the novel formal treatment in the theories of Dependency and Government Phonology (GP, e.g. Kaye et al. (1990), Harris (1994)). These latter approaches posit the so called primes, or basic elements, that are monovalent (c.f. binary SPE features). For example, there are four basic resonance primes commonly labeled as A, U, I, and @; the first three denoting the peripheral vowels [a], [u] and [i] respectively, the last one describing the most central vowel quality of schwa. English [iː] would correspond to the I prime while [e] results in fusing the I and A primes. In addition to these 'vocalic' primes, GP proposes the 'consonantal' primes ?, h, H, N, denoting closure, friction, voicelessness and nasality respectively. To account for more varied inventories (e.g. English has 20 contrastive vowels and thus 20 different phonetic representations for them), phonetic qualities and phonological behaviour, the phoneme representations based on primes become heavily perplexed. Most importantly, and given the name of the framework, some elements can optionally be heads and govern the presence or realisation of other (dependent) elements.

These developments established the relevance of the internal structure of phonological primitives and their non-linear, and hierarchical nature. Importantly, while the SPE-style features were assumed to require the full interpretation of all features for a surface phonetic realisation of a phoneme, the primes of GP are assumed to be interpretable alone despite their sub-phonemic nature. Harris and Lindsey (1995) call this GP assumption the autonomous interpretation hypothesis.

This hypothesis, and its testing, is at the core of our approach. One of the goals of this work is to provide interpretation grounded in the acoustic signal for sub-phonemic representational components of phonology.

2.2 Representing CMUbet with SPE and GP

To characterise the phoneme inventory of American English in the SPE and the GP frameworks, we have adopted the approach of King and Taylor (2000) with some modifications. We use the reduced set of 39 phonemes in the CMUbet system[1]. The major difference regarding the SPE-style representations is our addition of [rising] feature used to differentiate diphthongs from monophthongs. In the original SPE framework, this difference was treated with the [long] feature and the surface representation of diphthongs was derived from long monophthongs through a rule. Given the absence of the 'rule module' in our approach to synthesis, we opted for a unified feature specification of the full vowel inventory of American English using the added feature. This allowed for diphthongs to form a natural class with vowels rather than with glides [j, w] as in King and Taylor (2000). Additional minor adjustments assured the uniqueness of a feature matrix for each phoneme. The full specification of all 39 phonemes with 14 binary features is shown in Tab. 8.

The set of GP-style specification for English phonemes can be seen in Tab. 7. Again, we followed King & Taylor, mostly in formalizing headedness with pseudo-features, which allows for GP phoneme specifications that are comparable to SPE. Especially the vowel specifications differ somewhat from King & Taylor stemming from our effort to approximate the phonetic characteristics of the vowels, and the differences among them, as closely as possible. For example, if the back lax [ʊ] is specified with E, the same was employed for the front lax [ɪ], or the front quality of [æ] was formalized with the A heading I compared to King & Taylor's formalism with only non-headed A.

3 Phonological vocoding

Both SPE and GP phonological systems represent a phone by a combination of phonological features. For example, a consonant [j] is articulated using the mediodorsal part of the tongue [+Dorsal], in the motionless, mediopalatal, part of the vocal tract [+High], generated with simultaneous vocal fold vibration [+Voiced]. These three classes then comprise the phonological representation for the [j] in the SPE system.

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

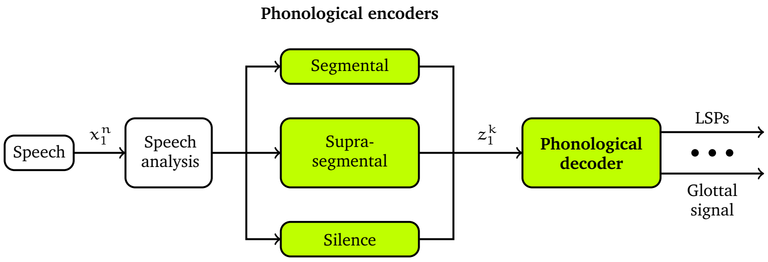


Figure 1: The process of the phonological vocoding. The encoder runs individual phonological encoders and merges phonological posteriors. The DNN-based decoder generates speech parameters – spectra lines LSPs and source parameters for speech re-synthesis.

A phonological vocoder is a cascaded DNN-based encoder and decoder that uses phonological features as speech parameters. Cernak et al. (2015) have recently proposed the vocoding for a low bit rate parametric speech coding. Fig. 1 shows a generalised process of the phonological vocoding.

3.1 Encoding

The encoder is based on a bank of phonological encoders realised as neural network classifiers that output K phonology posterior features $z_1^k = p(C_k|x_1^n)$, for $k = 1, \dots, K$, where C_k are individual natural classes of the phonological features. The posteriors are z_1^k can be optionally pruned (compressed) or manipulated. Manipulation of the encoded phonological features allows real-time synthesis of arbitrary speech sounds of the target voice.

We identified two main groups of the phonological features:

- segmental: phonological features that define a phonetic surface of words,
- supra-segmental: phonological features at a timescale of stress/accent and syllable.

3.2 Decoding

The phonological decoder is based on a DNN that learns the highly-complex regression problem of mapping posteriors z_1^k to speech parameters for re-synthesis. While phonological encoders are speaker-independent, the phonological decoder is speaker-dependent because of speaker-dependent speech parameters. The DNN is trained from available free audio-books.

The speech parameters are smoothed using dynamic features and pre-computed (global) variances Tokuda et al. (1995), and the formant enhancement Ling et al. (2006) is performed to compensate over-smoothing of the formant frequencies. Speech is re-synthesised using LPC re-synthesis with minimum-phase complex cepstrum glottal model estimation (Garner et al., 2015), from synthesized LSPs p_i, and source signal $\hat{\theta}$ and magnitude $\log(\hat{m})$ parameters, as shown by Eq. 1:

$$\hat{x}_n = \underbrace{\sum_{i=1}^{P} h(p_i|\hat{z}_1^k(n))\hat{x}(n-i)}_{\text{spectra}} + \underbrace{h(\theta_n, m_n, r_n|\hat{z}_1^k(n))}_{\text{source}}, \tag{1}$$

where $h(.)$ denotes nonlinear activation function (forward propagation) of the trained phonological decoder, and $\hat{z}_1^k(n)$ are binary phonological posteriors for time n.

4 Experimental setup for laboratory phonology

4.1 Phonological analysis – encoding

The encoder is based on a bank of phonological encoders realised as neural network classifiers – multilayer perceptrons (MLPs) – that encode individual phonology features. Each MLP classifies a binary phonological feature. We trained HMM/GMM systems for phonetic encoding using Wall Street Journal WSJ0 and WSJ1 continuous speech recognition corpora (Paul and Baker, 1992). The phoneme set comprising of 40 phonemes (including "sil", representing silence) was defined by the CMU pronunciation dictionary.

Table 1: *Classification accuracies (%) of the GP prime detectors at frame level.*

Prime	Accuracy (%)		Prime	Accuracy (%)	
	train	cv		train	cv
A	92.3	91.7	a	97.9	97.6
I	94.9	94.6	i	96.1	96.4
U	94.4	94.1	u	97.5	97.7
E	92.6	91.9	H	95.4	95.0
S	95.2	94.7	N	98.2	98.1
h	95.9	95.4	sil	99.0	98.9

Table 2: *Classification accuracies (%) of the SPE detectors at frame level.*

Natural classes	Accuracy (%)		Natural classes	Accuracy (%)	
	train	cv		train	cv
vocalic	96.0	95.8	round	97.8	97.7
consonantal	94.5	93.9	tense	94.8	94.1
high	94.8	94.4	voice	94.6	94.4
back	93.9	93.4	continuant	95.6	95.2
low	97.4	97.1	nasal	98.1	98.0
anterior	94.4	94.0	strident	97.8	97.6
coronal	94.3	93.9	sil	99.0	98.9

First, we trained an HMM/GMM system using PLP features. The three-state, cross-word triphone models were trained with the HTS variant of Zen et al. (2007) of the HTK toolkit on the *si_tr_s_284* set of 37,514 utterances. We tied triphone models with decision tree state clustering based on the minimum description length (MDL) criterion (Shinoda and Watanabe, 1997). The MDL criterion allows an unsupervised determination of the number of states. In this study, we obtained 12,685 states and modeled each state with a GMM consisting of 16 Gaussians.

Then, a bootstrapping phoneme alignment was obtained using forced alignment with cross-word triphones. The bootstrapping alignment was used for the training of 3-hidden layer 2000x500x2000 MLP, using temporal context of 9 successive frames of PLP features, and softmax output function. The architecture of the MLP was determined empirically. Using a hybrid HMM/MLP speech decoder fed with the phoneme posteriors, the re-alignment was performed. After two iterations of the MLP trainings and re-alignments, the best phoneme alignment of the speech data was obtained. This re-alignment increased the cross-validation accuracy of the MLP training from 77.54% to 82.36%. The representation of A was used to map the phonemes of the best alignment for the training of the encoders. Each encoder was trained with the frame alignment having two output labels, the encoded class present or not. The encoding MLPs were then trained again with the same settings as the alignment MLP training.

Tab. 1, Tab. 2, and Tab. 3 show classification accuracy at frame level of the GP, SPE and eSPE encoders, respectively. The encoder performances are high, with an average cross-validation training accuracy of 95.5%, 95.6% and 96.3%, respectively.

4.2 Phonological synthesis – decoding

The DNN of the phonological decoder converts the phonological posteriors z_1^k to the speech parameters. Generation of a particular speech sound requires to combine the associated phonological features at the DNN input or output. In the first case, we refer to this generation as **network-based**. The speech parameters are smoothed using dynamic features and pre-computed (global) variances, and the formant enhancement is performed with $\alpha = 0.5$ to compensate over-smoothing of the formant frequencies.

In the second case, we artificially provide a single phonological feature as an active input and the rest as zeros, generating a template of the speech parameters that characterise the input phonological feature. Then the speech sound is generated by the combination (audio mixing) of the outputs of the DNN. We refer to this case as **signal-based**. We wanted to test the suitability of speech sound synthesis without the DNN, i.e., only by mixing the artificial phonological sound components, and in addition, to allow the reader to experiment with the sound components that are embedded in this manuscript.

Table 3: *Classification accuracies (%) of the eSPE detectors at frame level.*

Natural classes	Accuracy (%)		Natural classes	Accuracy (%)	
	train	cv		train	cv
vowel	94.7	94.3	low	97.5	97.2
fricative	97.3	97.0	mid	94.5	94.0
nasal	98.2	98.1	retroflex	98.6	98.4
stop	96.7	96.4	velar	98.9	98.8
approximant	97.2	96.9	anterior	94.8	94.3
coronal	94.8	94.4	back	94.2	93.8
high	94.6	94.2	continuant	95.8	98.4
dental	99.3	99.2	round	95.3	94.9
glottal	99.7	99.7	tense	91.4	90.7
labial	97.6	97.4	voiced	95.2	94.9

4.2.1 Training

For training the DNN of the phonological decoder we selected an English audio book "Anna Karenina" of Leo TOLSTOY[2], around 36 hours long. Recordings were organised into 238 sections, and we used the sections 1–209 as a training set, 210–230 as a development (cross-validation) set and 231–238 as a testing set. The development and testing sets were 3 hours and 1 hour long, respectively.

The speech signals sampled at 16 kHz, framed by 25-ms windows with 10-ms frame shift, were used for extraction of the following speech parameters:

- static Line Spectral Pairs (LSPs) of 24th order,
- gain $\log(g)$, continuous pitch $\log(F0)$,
- a Harmonic-To-Noise (HNR) ratio $\log(r)$,
- and two glottal model parameters – angle θ and magnitude $\log(m)$ of a glottal pole.

Extraction was done by the Speech Signal Processing (SSP) python toolkit[3]. We used static speech parametrization of 29th order along with its dynamic features, altogether of 87th order as DNN output features.

Phonological posteriors z_1^k were used as the DNN input features. They were extracted with the same 10-ms frame shift, and thus almost perfect frame alignment was obtained. Temporal context of 11 successive frames resulted into the input feature vector of 264 dimensions. Cepstral mean and variance normalisation of the input features was applied before the training.

The DNN was initialised using 4x1024 Deep Belief Network pre-training by contrastive divergence with 1 sampling step (CD1) (Hinton et al., 2006). The DNN with a linear output function was then trained using a mini-batch based stochastic gradient descent algorithm with mean square error cost function of the KALDI toolkit (Povey et al., 2011). The DNN was trained with 3.4 million parameters.

5 Application examples for laboratory phonology

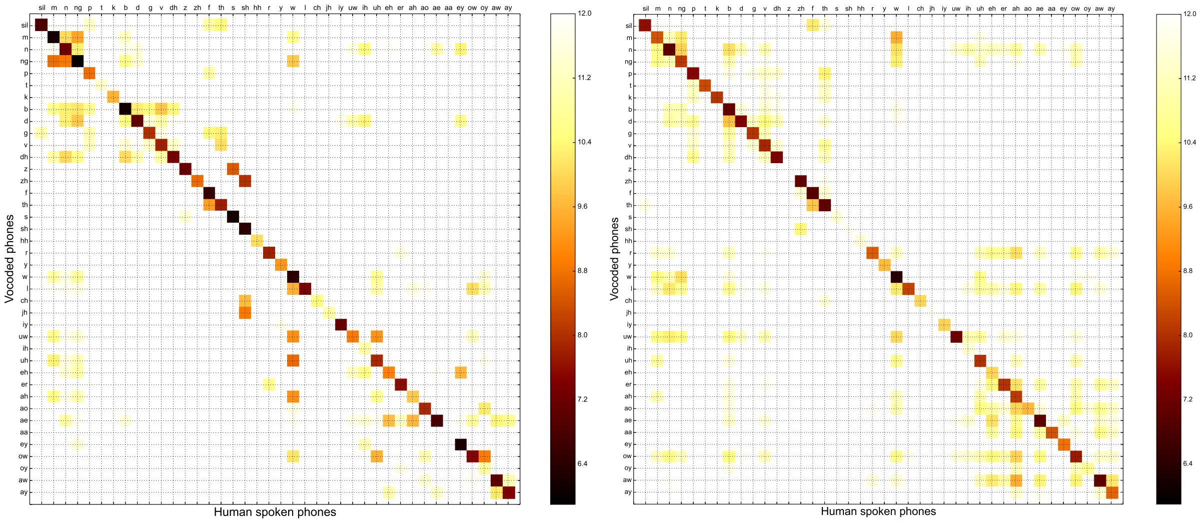
In this section we show how to use the phonological vocoders described in Section 4 for comparison of the phonological systems. We then continue with the second example of the parametric phonological phoneme-to-speech synthesis.

5.1 Comparison of the phonological systems

We start with the context-independent vocoding in Section 5.1.1, i.e., the vocoding of isolated speech sounds, and the context-dependent vocoding in Section 5.1.2.

[2] https://librivox.org/anna-karenina-by-leo-tolstoy-2/

[3] https://github.com/idiap/ssp



(a) Network-generated

(b) Signal-combined

Figure 2: *MCD distances between GP vocoded phones and human spoken phones.*

5.1.1 Context-independent vocoding

In the context-independent vocoding, we generated oracle phonological posteriors, i.e., using only the features that represent the specific isolated phone (rows of Tab. 8, 7 and 9). We were interested in the evaluation of the confusions among the phones produced by the phonological vocoding and human phones. For the human phones, we manually phonetically labelled 76 utterances from the audio book test set. Having the phone boundaries, we concatenated all central frames (trying to eliminate the co-articulation) of the same phonemes to obtain human spoken acoustic templates.

We calculated the Mel Cepstral Distortion (MCD) (Kubichek, 1993) between the vocoded phones and human-spoken phones. There is an intrinsic phoneme confusion in human recognition and as we were interested only in those confusions introduced by the phonological representation, we normalised the calculated distortions as follows:

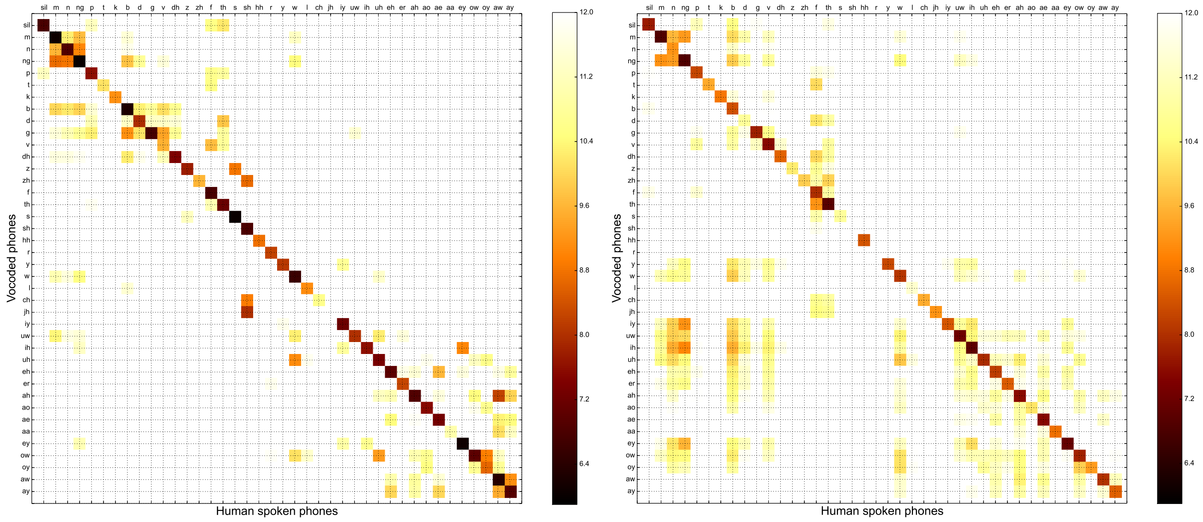
$$C_{norm} = C * (1 + \mathrm{scale}(H)/100) \tag{2}$$

$$\mathrm{scale}(H) = 100 - 100 * \frac{\max(H) - H}{\max(H) - \min(H)} \tag{3}$$

where C stands for confusions between vocoded and human phones and H stands for confusions between human phones only. The MCD gives higher numbers for more confused phones. The scaling defined by Eq. 2 represents normalised human phoneme confusion, giving zero for the same sounds. The human confusions are thus removed from the vocoded ones, and we could evaluate the contributions of the phonological features to the vocoded phone confusions.

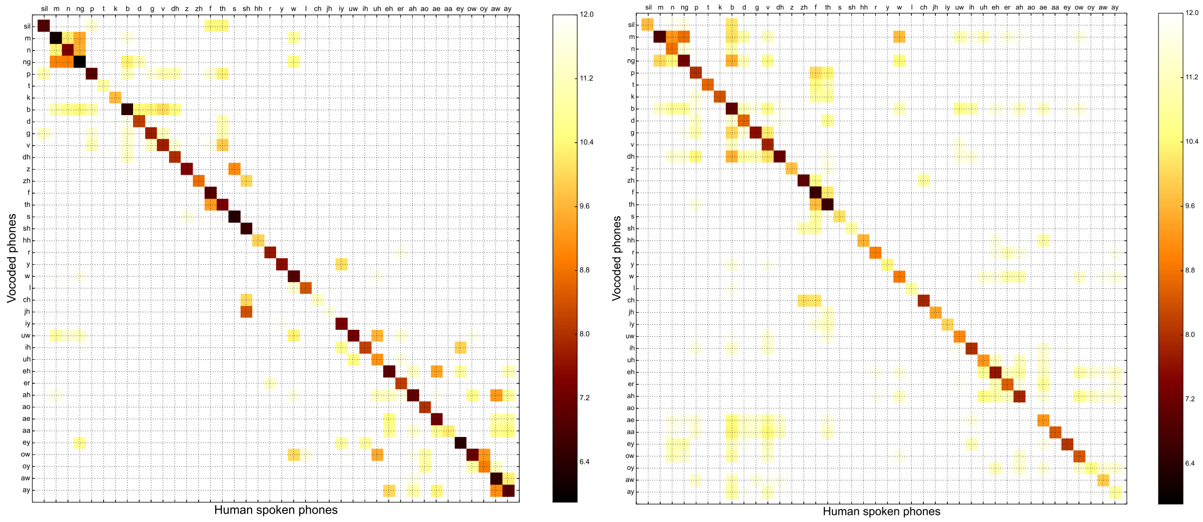
Fig. 2,3,4 show normalised confusions C_{norm} of vocoded context-independent phonemes and human-produced phonemes of the same speaker, for the GP, SPE and eSPE phonological system, respectively. The diagonal elements of the confusion matrices represent an acoustical distance (dissimilarity) of vocoded and human phones. If the phonological features represent speech well, the matrices show only strong diagonals. The missing diagonal values or higher confusions between phones imply errors of the phonological speech representation, usually wrongly assigned phonological features to the phone.

The figures show two aspects of the evaluation. In the first, the confusion matrices of (a) the network-based and (b) the signal-based phonological decoding are shown. The network- and signal-based decoding differ in the combination of the DNN inputs and the outputs, respectively. We can therefore consider this as two different evaluation metrics, hypothesising that both partially contribute to a final evaluation. For all the three phonological systems, the results of signal-based context-independent vocoding show higher errors than the ones of network-based vocoding. This might be caused by a fact that audio mixing at DNN output is a linear operation (as shown in following Section 5.2.1), and it is an approximation to a non-linear function that is modelled by network-based phonological decoding. Additionally, the types of reported confusions in the network-based vocoding that are missing in the signal-based one make sense



(a) Network-generated (b) Signal-combined

Figure 3: *MCD distances between SPE vocoded phones and human spoken phones.*



(a) Network-generated (b) Signal-combined

Figure 4: *MCD distances between eSPE vocoded phones and human spoken phones.*

phonetically. For example, in the left panel of Fig. 2a, phonetically very similar [ʒ], [ʃ], [dʒ], [tʃ] with voicing and closure being highly context dependent are confused. Nevertheless, signal-based vocoding tends to produce 'tighter' confusions in all three frameworks. For example, while [ð] in the left panel of Fig. 2a shows confusions with nasals, voiced plosives and [v], it shows only minor confusions labials with in the right panel.

In the second aspect of the evaluation, comparing Fig. 2a, Fig. 3a, and Fig. 4a, we see different error patterns. In all three phonological systems the biggest confusions are shown with the nasals [m n ŋ]. GP in addition produces confusions of the consonants and vowels, such as for nasals. SPE seems to represent speech better, namely for the vowel [ɑ] and glide [w], suppressing most of the vowel-consonant confusions. On the other hand, it fails with proper [dʒ] vocoding. We found that this phone was the less frequent in our evaluation data that may cause these less significant measurements. Finally, according to our data, the eSPE further improves the SPE speech representation. It generates less confusions in the vowel space, and also in the consonant space, such as for the voiced stops [b d ɡ].

5.1.2 Context-dependent vocoding

The previous experiment evaluated the vocoding of the isolated sounds. We continued with the evaluation of continuous speech. Here, signal-based phonological decoding was not used.

To evaluate continuous phonological speech representations, we vocoded all the test set of the audio book, and subjectively compared the systems. We were interested if the segmental errors found in context-independent vocoding impact the context-dependent vocoding.

An ABX subjective evaluation listening test was performed between the GP-based and the eSPE-based vocoded speech, the systems with the highest differences in context-independent results. 15 listeners, expert and non-experts in speech processing participated in the listening test. The subjects were asked to listen to sample pairs of sentences (as many times as they wanted), and choose between the two paired samples in terms of their overall quality. The samples were presented to them in a random order with no indication of which system they represented. Additionally, the listeners could choose a third option, *both samples sound the same*, if they had no preference between them. The same 8 sentences from the test set were used for all listeners, which produced a total of 16 samples.

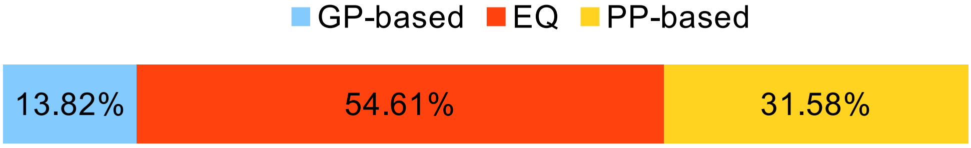


Figure 5: ABX subjective evaluation listening test between the GP-based and the eSPE-based vocoded speech

In Fig. 5 the results of the ABX test for the GP and eSPE phonological systems, are shown. As can be seen, the eSPE-based phonological vocoder outperforms the GP one by approximately 32% over 14% preference score in the ABX test. Even though there is a preference of the listeners towards the eSPE-based system, nevertheless, it is clearly shown that with a very high percentage, almost 55%, the two systems are perceived by the listeners as having the same overall quality. It should be pointed out that a t-test confirmed that the differences between the GP-based and eSPE-based phonological vocoders are statistical significant ($p < 0.01$). Subsequent auditory and visual analysis of the sample sentences and their generated acoustic signals suggests that eSPE sentences tended to be produced with slightly greater articulatory effort, and occasional preference for American English over British English. More effort was acoustically realized as longer closures for plosives, stronger releases, and also slightly greater disjunctures at some word boundaries. All of these could contribute to a clearer perception of the eSPE sentences.

The vocoded continuous speech was fully intelligible, and we conclude that the phonological vocoder certainly learns some supra-segmental information (using the temporal window of 11 successive frames – around 100 ms of speech) that may correspond to the formant transitions and differences in voice onset times. This is probably enough to learn co-articulation well.

5.2 Experimental parametric phonological TTS

5.2.1 Isolated phone synthesis

We hypothesise that acoustic representation of the phonological features forms a set of acoustic templates that define whole speech acoustic space similarly as their phonological counterparts define the phonetic surface of a language. In this section we show how these acoustic templates could be combined to generate arbitrary speech sounds, and how they might be used for speech synthesis from canonical phone representation. In GP, speech sounds are created by fusion and splitting of primes, so we selected the GP system for further experiments.

In GP, speech sounds are created by the primes, the sound pattern templates, by reference to which speakers orchestrate articulatory output (Harris, 1994). Here we illustrate how the acoustic representation of these primes can be generated within our platform. We call these acoustic realizations **sounds of primes**.

Given our hypothesis mentioned above (that these acoustic representations x_s form a set of acoustic templates that define the speech acoustic space), we created artificial phonological sound representation z_1^S according to Eq. 4, repeated patterns to get sounds x_s two seconds long.

$$z_s = \begin{cases} 1 & \text{if } s = k, \text{ where } k \in 1, 2, \ldots, K \\ 0 & \text{if } s \neq k, \text{ where } k \in 1, 2, \ldots, K \end{cases} \tag{4}$$

Table 4: *Recordings demonstrating individual sounds of primes, 2 seconds long.*

Prime	Sound of prime	Prime	Sound of prime
A	(wav)	a	(wav)
I	(wav)	i	(wav)
U	(wav)	u	(wav)
H	(wav)	h	(wav)
S	(wav)	N	(wav)
E	(wav)		

Tab. 4 demonstrates recordings[4] of individual sounds of primes. Harris (1994) claims that fusing and splitting of primes accounts for phonological description of the sound. We followed the Harris's phonological rule (29)[5], and tried to synthesise non-English sounds by fusing involved primes. We claim than new sounds can be generated by simple mixing of the corresponding acoustical primes x_s. Tab. 5 demonstrates synthesis of standard German sounds [y] and [œ] from English x_s, generated as follows:

$$x_m = \frac{1}{S} \sum_{s=1}^{S} w_s x_s, \tag{5}$$

where x_m is a fusion of S acoustical primes x_s, and w_s are weights of the fusion. For $w_s = 1$, it can be done easily with available free tools, e.g.:

```
[y ] : sox -m I.wav U.wav E.wav y.wav
[œ ] : sox -m A.wav I.wav U.wav E.wav oe.wav
```

The fusion of Eq. 5 represents a static mixing of S acoustical primes, i.e., it cannot be applied to model co-articulation. To include co-articulation into the synthesis, the phonological decoder has to be used. As it was trained with the temporal context of 11 successive frames, around 50ms before and 50ms after current processing frame, it learnt how speech parameters change with trajectories of the phonological posteriors.

Table 5: *Recordings demonstrating fusing of primes, resulting into the synthesis of new sounds.*

Combination	IPA	Sound of prime
[A, I, U, E]	œ	(wav)
[I, U, E]	y	(wav)

Apparently, the signal-based phonological decoding works well not only for the GP phonological system, but also for the SPE and eSPE phonological systems, as demonstrated in Section 5.1.1.

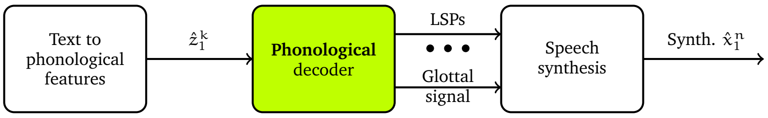


Figure 6: The speech synthesis with the phonological decoding. Speech parameters are generated by the DNN, the speech spectra lines LSPs and source parameters. Speech is generated by subsequent LPC re-synthesis.

[4] The recordings are downloadable

[5] [I, U, E] → y; [A, I, U, E] → œ

5.2.2 Continuous speech synthesis

Experimental parametric phonological TTS is designed by a simplistic text processing front end: text → phonemes → phonological features. Fig. 6 shows the speech synthesis process with the phonological decoding. The binary phonological representation to be synthesized $\hat{z}_1^k$ is obtained from the canonical phone representation.

To demonstrate the parametric phonological TTS system, we selected a female slt testing speaker from the CMU-ARCTIC speech database (Kominek and Black, 2004). We took the phoneme symbols along with their durations, and mapped it to the phonological representation. Then we synthesised the sentences using the phonological decoder.

Tab. 6 demonstrates recordings of speech synthesis from the phonological speech representation. The example slt_cmu_us_arctic_a0453 demonstrates how the phonological vocoder learns the context. The phoneme sequence of the first word is [eɪ t i n], while the synthesised sequences using both phonological systems are rather [eɪ tʃ i n]. It illustrates assimilation of place or articulation of the synthesised phoneme [t]. It takes the place of articulation of the following phoneme [i], the high portion of the tongue, that results into affricate [tʃ]. Additionally, since [t] begins a stressed syllable, it is aspirated. The acoustic quality of the release burst and the subsequent interval of aspiration are similar to lingually created frication of sibilants due to the high position of the tongue.

Table 6: *Recordings demonstrating phonological speech synthesis.*

Sentence	GP	SPE	eSPE
slt_cmu_us_arctic_a0453	(wav)	(wav)	(wav)
slt_cmu_us_arctic_a0457	(wav)	(wav)	(wav)
slt_cmu_us_arctic_a0460	(wav)	(wav)	(wav)

For evaluating the phonological TTS, an informal intelligibility test was conducted using semantically unpredictable sentences (SUS). Seven SUSs, taken from SIWIS project – Spoken Interaction with Interpretation in Switzerland (SIWIS)[6], were synthesised. The length of the sentences varied from 6 to 8 words. Four native and non-native English speakers, experts in the speech processing field, listened to each synthesized sentence (less than three times each) and transcribed the audio. As a reference for the intelligibility test, the same sentences were synthesized by a hidden Markov model (HMM) parametric speech synthesizer.

For building the HMM models, the HTS v. 2.1 toolkit (HTS, 2010) was used. Specifically, the implementation from the EMIME project (Wester et al., 2010) was taken. The speech data which were used had 16kHz sampling frequency. Five-state, left-to-right, no-skip HSMMs were used. The speech parameters which were used for training the HSMMs were 39 order mel-cepstral coefficients, log-F0 and 21-band aperiodicities, along with their delta and delta-delta features, extracted every 5 ms. For training the HMM models, the same training set of the audio book which was used for training the phonological vocoder, was used.

The average intelligibility score achieved synthesizing speech using the phonological TTS was 58% of correct words in comparison to the HMM-based TTS where the listeners achieved an average of 81% of correct words. The phonological TTS thus achieves intelligibility of 71% of the state-of-the-art parametric TTS. Since the voice for the phonological TTS is built in an unsupervised manner from the audio book, and without any TTS front-end, we consider these results promising.

6 Conclusions

We have proposed to use speech vocoding as a platform for laboratory phonology. The proposal consists of a cascaded phonological analysis and synthesis. The objective and subjective evaluations supported the hypothesis that the most informative feature set achieves the best quality vocoded speech, where the features are verified in both directions, recognition and synthesis, simultaneously.

We have showed two application examples of our proposed approach. First, we compared three systems of phonological representations and concluded that eSPE achieves slightly better results than the other two. Our results thus support other recent work showing that eSPE is suitable for phonological analysis, for speech recognition and language identification tasks (Yu et al., 2012; Siniscalchi et al., 2012).

[6] http://www.idiap.ch/project/siwis

However, GP – the most compact phonological speech representation, performes in the analysis/synthesis tasks comparably to the systems with higher number of phonological classes.

Second, we presented the synthesis of acoustic representation of the phonological features (sounds of primes), and proposed a model for generating arbitrary speech sounds. In addition, we explored phonological parametric TTS without any front-end, trained from an unlabelled audio book in an unsupervised manner, and achieving intelligibility of 71% of the state-of-the-art parametric speech synthesis. This seems to be a promising approach for unsupervised and multilingual text-to-speech systems.

In this work we have focused on segmenal evaluation of the phonological system. In the future, we plan to incorporate supra-segmental features, and use the proposed laboratory phonology platform for further experimentation such as generation of speech stimuli for perception experiments.

We envision that the presented approach paves the way for researchers in both fields to form meaningful hypotheses that are explicitly testable using the concepts developed and exemplified in this paper. Laboratory phonologists might test the compactness, confusability, perceptual viability, and other applied concepts of their theoretical models. This might be done in at least two ways. First, synthesis/recognition tests might follow hand in hand with their analysis of data from human speech production and perception, which would allow for accumulating much needed data on the differences between human and machine performance. Second, synthesis/recognition might be used for pre-testing before undergoing experiments and analysis of human data, which is commonly time and effort demanding. Speech processing community may be suggesting tests for the potential of advancing the performance of current state-of-the-art applications, for example in multi-lingual processing, using the concepts developed for the theoretical phonological models. If these concepts formulated to be testable with the proposed platform, the chances are that they are readily transferable to speech processing.

7 Acknowledgements

We would like to Xingyu Na from Institute of Acoustics, Chinese Academy of Sciences, for his valuable comments on an early draft of this paper.

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS)

References

Thomas G. Bever and David Poeppel. Analysis by Synthesis: A (Re-)Emerging Program of Research for Language and Vision. *Biolinguistics*, 4(2-3):174–200, 2010.

Milos Cernak, Blaise Potard, and Philip N. Garner. Phonological vocoding using artificial neural networks. In *Proc. of ICASSP*. IEEE, April 2015. URL https://publidiap.idiap.ch/index.php/publications/show/3070.

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

G. N. Clements and E. Hume. *The Internal Organization of Speech Sounds*, pages 245–306. Oxford: Basil Blackwell, Oxford, 1995.

Philip N. Garner, Milos Cernak, and Blaise Potard. A simple continuous excitation model for parametric vocoding. Technical Report Idiap-RR-03-2015, Idiap, January 2015. URL http://publications.idiap.ch/index.php/publications/show/2955.

Anne-Lise L. Giraud and David Poeppel. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4):511–517, April 2012. ISSN 1546-1726. URL http://view.ncbi.nlm.nih.gov/pubmed/22426255.

J. Goldsmith. *Autosegmental Phonology*. PhD thesis, MIT, Cambridge, Massachusetts, 1976.

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

John Harris. *English Sound Structure*. Wiley-Blackwell, 1 edition, December 1994. ISBN 0631187413. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0631187413.

Geoffrey E. Hinton, Simon Osindero, and Yee W. Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.*, 18(7):1527–1554, July 2006. ISSN 0899-7667. doi: 10.1162/neco.2006.18.7.1527. URL http://dx.doi.org/10.1162/neco.2006.18.7.1527.

D. Hirst. The analysis by synthesis of speech melody: from data to models. *Journal of Speech Sciences*, 1 (1):55–83, 2011.

HTS. HMM-based speech synthesis system version 2.1. 2010. URL http://hts.sp.nitech.ac.jp.

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

J. Kaye, J. Lowenstamm, and Jean-Roger Vergnaud. Constituent structure and government in phonology. *Phonology*, 7(2):193–231, 1990.

Simon King and Paul Taylor. Detection of phonological features in continuous speech using neural networks. *Computer Speech & Language*, 14(4):333–353, October 2000. ISSN 08852308. doi: 10.1006/csla.2000.0148. URL http://dx.doi.org/10.1006/csla.2000.0148.

J. Kominek and A. Black. The CMU Arctic speech databases. In *Proc. of 5th ISCA Speech Synthesis Workshop*, pages 223 – 224, 2004.

R. F. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proc. of ICASSP*, volume 1, pages 125–128 vol.1. IEEE, May 1993. ISBN 0-7803-0971-5. doi: 10.1109/pacrim.1993.407206. URL http://dx.doi.org/10.1109/pacrim.1993.407206.

W. Leben. *Suprasegmental Phonology*. PhD thesis, MIT, Cambridge, Massachusetts, 1973.

Ki-Seung Lee and R.V. Cox. A very low bit rate speech coder based on a recognition/synthesis paradigm. *IEEE Trans. on Audio, Speech, and Language Processing*, 9(5):482–491, Jul 2001. ISSN 1063-6676. doi: 10.1109/89.928913.

Zhen-Hua Ling, Yi-Jian Wu, Yu-Ping Wang, Long Qin, and Ren-Hua Wang. USTC system for Blizzard Challenge 2006 - an improved HMM-based speech synthesis method. In *Proc. of Blizzard Challenge workshop*, 2006.

Douglas B. Paul and Janet M. Baker. The design for the wall street journal-based CSR corpus. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, pages 357–362, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. ISBN 1-55860-272-0. doi: 10.3115/1075527.1075614. URL http://dx.doi.org/10.3115/1075527.1075614.

J. Picone and G. R. Doddington. A phonetic vocoder. In *Proc. of ICASSP*, pages 580–583 vol.1. IEEE, May 1989. doi: 10.1109/icassp.1989.266493. URL http://dx.doi.org/10.1109/icassp.1989.266493.

Janet B. Pierrehumbert, Mary E. Beckman, and D. Robert Ladd. *Conceptual foundations of phonology as a laboratory science*, pages 273–303. Oxford University Press, Oxford, 2000.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *Proc. of ASRU*. IEEE SPS, December 2011. IEEE Catalog No.: CFP11SRW-USB.

E. C. Sagey. *The representation of features and relations in non-linear phonology*. PhD thesis, MIT, Cambridge, Massachusetts, 1986.

Koichi Shinoda and Takao Watanabe. Acoustic modeling based on the MDL principle for speech recognition. In *Proc. of Eurospeech*, pages I –99–102, 1997.

S. M. Siniscalchi, Dau-Cheng Lyu, T. Svendsen, and Chin-Hui Lee. Experiments on Cross-Language Attribute Detection and Phone Recognition With Minimal Target-Specific Training Data. *IEEE Trans. on Audio, Speech, and Language Processing*, 20(3):875–887, March 2012. ISSN 1558-7916. doi: 10.1109/tasl.2011.2167610. URL http://dx.doi.org/10.1109/tasl.2011.2167610.

K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *Proc. of ICASSP*, volume 1, pages 660–663 vol.1. IEEE, May 1995. ISBN 0-7803-2431-5. doi: 10.1109/icassp.1995.479684. URL http://dx.doi.org/10.1109/icassp.1995.479684.

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*, volume 2, pages 609–612 vol.2. IEEE, May 1998. ISBN 0-7803-4428-6. doi: 10.1109/icassp.1998.675338. URL http://dx.doi.org/10.1109/icassp.1998.675338.

Mirjam Wester, John Dines, Matthew Gibson, Hui Liang, Yi-Jian Wu, Lakshmi Saheer, Simon King, Keiichiro Oura, Philip N. Garner, William Byrne, Yong Guan, Teemu Hirsimäki, Reima Karhila, Mikko Kurimo, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, and Junichi Yamagishi. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *SSW*, pages 192–197, 2010. URL http://www.isca-speech.org/archive/ssw7/ssw7_192.html.

Dong Yu, Sabato Siniscalchi, Li Deng, and Chin-Hui Lee. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In *Proc. of ICASSP*. IEEE SPS, March 2012. URL http://research.microsoft.com/apps/pubs/default.aspx?id=157585.

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

A Mapping of the phonological features to CMUbet

Following Tab. 7,8 and 9 show the mapping of phonological features to the used phonemes in this work.

Table 7: *GP phonological features and their association to CMUbet phonemes used in this paper.*

CMUbet	IPA	A	I	U	E	S	h	H	N	a	i	u	silence
iy	i	−	+	−	−	−	−	−	−	−	+	−	−
ih	ɪ	−	+	−	+	−	−	−	−	−	−	−	−
uw	u	−	−	+	−	−	−	−	−	−	−	+	−
uh	ʊ	−	−	+	+	−	−	−	−	−	−	−	−
ey	eɪ	+	+	−	−	−	−	−	−	−	+	−	−
ow	oʊ	+	−	+	−	−	−	−	−	−	−	+	−
oy	oɪ	+	−	+	−	−	−	−	−	−	+	+	−
ao	ɔ	+	−	+	+	−	−	−	−	−	−	+	−
aa	ɑ	+	−	−	−	−	−	−	−	+	−	−	−
ae	æ	+	+	−	−	−	−	−	−	+	−	−	−
ah	ʌ	+	−	−	+	−	−	−	−	−	−	−	−
aw	aʊ	+	−	+	−	−	−	−	−	−	−	+	−
ay	aɪ	+	+	−	−	−	−	−	−	−	+	−	−
y	j	−	+	−	−	−	−	−	−	−	−	−	−
w	w	−	−	+	−	−	−	−	−	−	−	−	−
eh	e	+	+	−	+	−	−	−	−	−	+	−	−
er	ɝ	+	−	−	+	−	−	−	−	−	−	−	−
r	ɹ	+	−	+	+	−	−	−	−	−	−	−	−
l	l	−	−	−	−	+	−	−	−	−	−	−	−
p	p	−	−	+	−	+	+	+	−	−	−	−	−
b	b	−	−	+	−	+	+	−	−	−	−	−	−
f	f	−	−	+	−	−	+	+	−	−	−	−	−
v	v	−	−	+	−	−	+	−	−	−	−	−	−
m	m	−	−	+	−	+	−	−	+	−	−	−	−
t	t	+	−	−	−	+	+	+	−	−	−	−	−
d	d	+	−	−	−	+	+	−	−	−	−	−	−
th	θ	+	−	−	−	−	+	+	−	−	−	−	−
dh	ð	+	−	−	−	−	+	−	−	−	−	−	−
n	n	−	−	−	−	+	−	−	+	−	−	−	−
s	s	−	−	−	+	−	+	+	−	−	−	−	−
z	z	−	−	−	+	−	+	−	−	−	−	−	−
ch	tʃ	−	+	−	−	+	−	+	−	−	−	−	−
jh	dʒ	−	+	−	−	+	−	−	−	−	−	−	−
sh	ʃ	−	+	−	−	−	+	+	−	−	−	−	−
zh	ʒ	−	+	−	−	−	+	−	−	−	−	−	−
k	k	−	−	−	+	+	+	+	−	−	−	−	−
g	ɡ	−	−	−	+	+	+	−	−	−	−	−	−
ng	ŋ	−	−	−	+	+	−	−	+	−	−	−	−
hh	h	−	−	−	−	−	+	+	−	−	−	−	−

Table 8: *SPE phonological features and their association to CMUbet phonemes used in this paper.*

CMUbet	IPA	vocalic	consonantal	high	back	low	anterior	coronal	round	rising	tense	voice	continuant	nasal	strident	silence
iy	i	+	−	+	−	−	−	−	−	−	+	+	+	−	−	−
ih	ɪ	+	−	+	−	−	−	−	−	−	−	+	+	−	−	−
uw	u	+	−	+	+	−	−	−	+	−	+	+	+	−	−	−
uh	ʊ	+	−	+	+	−	−	−	+	−	−	+	+	−	−	−
ey	eɪ	+	−	−	−	−	−	−	−	+	+	+	+	−	−	−
ow	oʊ	+	−	−	+	−	−	−	+	+	+	+	+	−	−	−
oy	oɪ	+	−	−	+	−	−	−	+	+	−	+	+	−	−	−
ao	ɔ	+	−	−	+	−	−	−	+	−	−	+	+	−	−	−
aa	ɑ	+	−	−	+	+	−	−	−	−	+	+	+	−	−	−
ae	æ	+	−	−	−	+	−	−	−	−	−	+	+	−	−	−
ah	ʌ	+	−	−	+	−	−	−	−	−	−	+	+	−	−	−
aw	aʊ	+	−	−	+	+	−	−	−	+	+	+	+	−	−	−
ay	aɪ	+	−	−	−	+	−	−	−	+	+	+	+	−	−	−
y	j	−	−	+	−	−	−	−	−	−	−	+	+	−	−	−
w	w	−	−	+	+	−	−	−	+	−	−	+	+	−	−	−
eh	e	+	−	−	−	−	−	−	−	−	−	+	+	−	−	−
er	ɝ	+	−	−	−	−	−	−	−	−	+	+	+	−	−	−
r	ɹ	+	+	−	−	−	−	+	−	−	−	+	+	−	−	−
l	l	+	+	−	−	−	+	+	−	−	−	+	+	−	−	−
p	p	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−
b	b	−	+	−	−	−	+	−	−	−	−	+	−	−	−	−
f	f	−	+	−	−	−	+	−	−	−	−	−	+	−	+	−
v	v	−	+	−	−	−	+	−	−	−	−	+	+	−	+	−
m	m	−	+	−	−	−	+	−	−	−	−	+	−	+	−	−
t	t	−	+	−	−	−	+	+	−	−	−	−	−	−	−	−
d	d	−	+	−	−	−	+	+	−	−	−	+	−	−	−	−
th	θ	−	+	−	−	−	+	+	−	−	−	−	+	−	−	−
dh	ð	−	+	−	−	−	+	+	−	−	−	+	+	−	−	−
n	n	−	+	−	−	−	+	+	−	−	−	+	−	+	−	−
s	s	−	+	−	−	−	+	+	−	−	−	−	+	−	+	−
z	z	−	+	−	−	−	+	+	−	−	−	+	+	−	+	−
ch	tʃ	−	+	+	−	−	−	+	−	−	−	−	−	−	+	−
jh	dʒ	−	+	+	−	−	−	+	−	−	−	+	−	−	+	−
sh	ʃ	−	+	+	−	−	−	+	−	−	−	−	+	−	+	−
zh	ʒ	−	+	+	−	−	−	+	−	−	−	+	+	−	+	−
k	k	−	+	+	+	−	−	−	−	−	−	−	−	−	−	−
g	g	−	+	+	+	−	−	−	−	−	−	+	−	−	−	−
ng	ŋ	−	+	+	+	−	−	−	−	−	−	+	−	+	−	−
hh	h	−	−	−	−	+	−	−	−	−	−	−	+	−	−	−

Table 9: *eSPE phonological features and their association to CMUbet phonemes used in this paper.*

CMUbet	IPA	vowel	fricative	nasal	stop	approxim.	coronal	high	dental	glottal	labial	low	mid	retroflex	velar	anterior	back	continuant	round	tense	voiced	silence
iy	i	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	+	−	+	+	−
ih	ɪ	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	+	−	−	+	−
uw	u	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	+	+	+	+	+	−
uh	ʊ	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	+	+	+	−	+	−
ey	eɪ	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−	+	+	−
ow	oʊ	+	−	−	−	−	−	+	−	−	−	−	+	−	−	−	+	+	+	+	+	−
oy	oɪ	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	−	+	−
ao	ɔ	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+	−
aa	ɑ	+	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	+	−	+	+	−
ae	æ	+	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	+	−	+	+	−
ah	ʌ	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−	+	+	−	−	+	−
aw	aʊ	+	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	+	+	+	+	−
ay	aɪ	+	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	+	−	+	+	−
y	j	−	−	−	−	+	−	+	−	−	−	−	−	−	−	−	−	+	+	−	+	−
w	w	−	−	−	−	+	−	−	−	−	+	−	−	−	−	+	−	+	+	−	+	−
eh	e	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−	−	+	−
er	ɝ	+	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	+	−
r	ɹ	−	−	−	−	+	−	−	−	−	−	−	−	+	−	−	−	+	+	−	+	−
l	l	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	+	−	−	+	−
p	p	−	−	−	+	−	−	−	−	−	+	−	−	−	−	+	−	−	−	+	−	−
b	b	−	−	−	+	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	+	−
f	f	−	+	−	−	−	−	−	−	−	+	−	−	−	−	+	−	+	−	+	−	−
v	v	−	+	−	−	−	−	−	−	−	+	−	−	−	−	+	−	+	+	−	+	−
m	m	−	−	+	−	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	+	−
t	t	−	−	−	+	−	+	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−
d	d	−	−	−	+	−	+	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−
th	θ	−	+	−	−	−	−	−	+	−	−	−	−	−	−	+	−	+	−	+	−	−
dh	ð	−	+	−	−	−	−	−	+	−	−	−	−	−	−	+	−	+	−	−	+	−
n	n	−	−	+	−	−	+	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−
s	s	−	+	−	−	−	+	−	−	−	−	−	−	−	−	+	−	+	−	+	−	−
z	z	−	+	−	−	−	+	−	−	−	−	−	−	−	−	+	−	+	−	−	+	−
ch	tʃ	−	+	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	+	−	−
jh	dʒ	−	+	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	+	−
sh	ʃ	−	+	−	−	−	−	+	−	−	−	−	−	−	−	−	−	+	−	+	−	−
zh	ʒ	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
k	k	−	−	−	+	−	−	+	−	−	−	−	−	−	+	−	+	−	−	+	−	−
g	ɡ	−	−	−	+	−	−	+	−	−	−	−	−	−	+	−	+	−	−	−	+	−
ng	ŋ	−	−	+	−	−	−	+	−	−	−	−	−	−	+	−	−	−	−	−	+	−
hh	h	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−