IDIAP RESEARCH REPORT

PHONOLOGICAL VOCODING USING ARTIFICIAL NEURAL NETWORKS

Milos Cernak Blaise Potard Philip N. Garner

Idiap-RR-04-2015

FEBRUARY 2015

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

Phonological vocoding using artificial neural networks

Milos Cernak, Blaise Potard, Philip N. Garner

February 4, 2015

Abstract

We investigate a vocoder based on artificial neural networks using a phonological speech representation. Speech decomposition is based on the phonological encoders, realised as neural network classifiers, that are trained for a particular language. The speech reconstruction process involves using a Deep Neural Network (DNN) to map phonological features posteriors to speech parameters – line spectra and glottal signal parameters – followed by LPC resynthesis. This DNN is trained on a target voice without transcriptions, in a semi-supervised manner. Both encoder and decoder are based on neural networks and thus the vocoding is achieved using a simple fast forward pass. An experiment with French vocoding and a target male voice trained on 21 hour long audio book is presented. An application of the phonological vocoder to low bit rate speech coding is shown, where transmitted phonological posteriors are pruned and quantized. The vocoder with scalar quantization operates at 1 kbps, with potential for lower bit-rate.

1 Introduction

We are interested in very low bit rate (VLBR) coding of speech. To this end, we have previously presented work on a phonetic vocoder Cernak et al. (2013a,b, 2014). The vocoder consists of automatic speech recognition (ASR) feeding into text to speech (TTS) synthesis. As only phonetic symbols need to be transmitted over a channel, this leads to a bit rate of only a few hundred bits per second (bps). Although the resulting speech quality is somewhat degraded, it is acceptable.

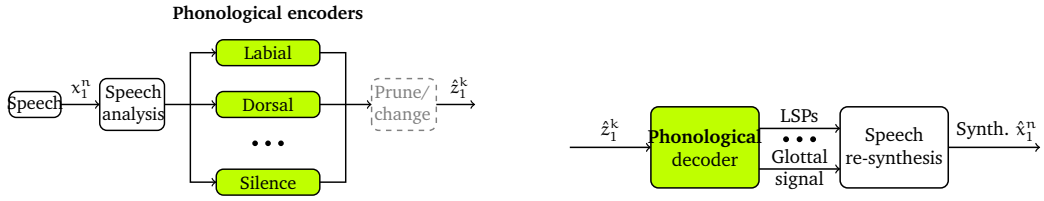
Recently, deep neural networks (DNNs) in hybrid HMM systems have led to large increases in ASR accuracy. At the same time, DNNs are also being used for TTS; results are very promising, e.g., Ze et al. (2013); Qian et al. (2014); Lu et al. (2013); Yin et al. (2014). In each case, much of the acoustic processing is taken on by the DNN, leaving the HMM to simply impose phonetic sequence constraints. This raises the question of whether the phonetic vocoder Picone and Doddington (1989) can benefit from such technology. Of course, the answer should be yes, in that the black-box functionality is the same. However, given the minimal contribution of the HMM, and that the application is simply vocoding, it also leads to the possibility of removing the HMM entirely.

The phonetic vocoder uses a phonetic representation of speech. Following the reasoning about using phonological features in speech processing of King and Taylor (2000), we hypothesise that the phonological representation is more suitable than the phonetic representation, because:

- The span of phonological features is wider than the span of phonetic features and thus the frame shift could be higher, i.e., fewer frames are transmitted yielding lower bit rates.
- The binary nature of phonological features promises to achieve a higher compression ratio.
- Phonological features are inherently multilingual Siniscalchi et al. (2012). This in turn has an attractive advantage in the context of multilingual vocoding without the need for a phonetic decision.

Phonological features can be categorised as distinctive features (primes) grouped into categories established from natural classes. There are several phonological systems such as Chomsky's system with binary features Chomsky and Halle (1968), multi-valued Ladefoged and Johnson (2014), and Government Phonology Harris and Lindsey (1995) feature systems. We followed a promising approach based on phone attributes (e.g, Yu et al. (2012); Siniscalchi et al. (2012)), and also used pseudo-phonological features in this work.

Fig. 1 sketches the design of the phonological vocoder. 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 $k = 1, \ldots, K$, where C_k are individual natural classes of the phonological features. The posteriors are



(a) Encoder (b) Decoder

Figure 1: The phonological vocoder split into (a) encoder and (b) decoder. The encoder runs individual phonological encoders and merges phonological posteriors. The decoder generates speech spectra lines LSPs and source parameters and re-synthesise the speech.

optionally pruned (compressed) or manipulated. The decoder is based on a DNN that learns a regression problem of mapping posteriors z_1^k to speech parameters used for re-synthesis.

We demonstrate the performance of the phonological vocoder on an experiment with French phonological encoders and the target voice trained on 20.9 hours male speech downloaded from a free public domain audiobook. A low bit rate speech coding achieved by pruning and quantizing of the phonological features is presented.

The structure of the paper is as follows: the next Section 2 introduces the French phonological features used in this work. Section 3 describes experiments and results on the training of the phonological vocoder, and applied to the parametric speech coding. Discussion and conclusions follow in section 4.

2 French phonological features

Phonological features can be used for speech sound classification Bauman-Waengler (2011). For example, a consonant [j] is articulated using the mediodorsal part of the tongue [+Dorsal class], in the motionless, mediopalatal, part of the vocal tract [+High class], generated with simultaneous vocal fold vibration [+Voiced class].

For natural classes of French sounds, we started from pseudo-phonological feature classification designed for American English Yu et al. (2012). We deleted the glottal and dental classes consisting of English phonemes [h, ð, θ], replaced [+Retroflex] with [+Uvular] consisting of a French rhotic consonant, and replaced the broad classes [+Continuant, +Tense] with:

- *Fortis* and *Lennis*, as an alternative to [+Tense] class, to distinguish consonants produced with greater and lesser energy, or articulation strength.
- *Alveolar* and *Postalveolar*, to distinguish between sibilants articulated by anterior portion of the tongue,
- *Dorsal*, to group consonants articulated by the central and posterior portions of the tongue,
- *Central*, to group vowels in the central position of the portion of tongue that is involved in the articulation and to the tongue's position relative to the palate Bauman-Waengler (2011),
- *Unround*, to group vowels with an opposite degree of lip rounding to the [+Round] class.

Tab. 1 shows the map of phonological features used in this study.

3 Experiments and results

3.1 Phonological encoders

The encoder is based on a bank of phonological encoders realised as neural network classifiers, 3-hidden layer multilayer perceptrons (MLPs), that encode individual phonology features. Each MLP classifies a binary phonological feature. The French speech database Ester Galliano et al. (2006) of standard French radio broadcast news was used for training of the encoders. It comprised 120 speakers in various recording conditions. We hypothesised that the broadcast recordings are more suitable for "live" speech encoding. In this study, a subset of 112 hours of recordings was used for the training. The phoneme set comprising 38 phonemes (including "sil") was defined by the BDLex Perennou (1986) lexicon.

Table 1: *French phonological features and their association to phonemes used in this paper, grouped into an organ, place and manner of articulation, and the others.*

	Class C_k	Phonemes
organ	Labial	b f m p v w ɥ
	Dorsal	ɲ ŋ j k g ʁ
	Coronal	d l n s t z ʃ ʒ
place	Alveolar	s z
	Postalveolar	ʃ ʒ
	High	y i ʃ u j g k ŋ
	Low	ɑ̃ a œ̃ ɔ o õ
	Mid	ø ẽ e œ ɛ
	Uvular	ʁ
	Velar	g k ŋ ɲ
manner	Vowel	i y u e ẽ ø o õ ə ɛ œ ɔ a ɑ̃ œ̃ ɐ
	Fricative	f v s ʃ z ʒ ʁ
	Nasal	m n ɲ ŋ ɑ̃ õ œ̃ ẽ
	Stop	b d g p t k
	Approximant	w l j ɥ
others	Anterior	b d f l m n p s t v z w
	Back	o õ ɔ u g k
	Lennis	b d g v z ʒ
	Fortis	f t p k s ʃ
	Round	o ɔ õ u œ œ̃ y ø
	Unround	a ɑ̃ i e ẽ ɛ ɐ
	Voiced	a ɑ̃ ɐ i y u e ẽ ø o õ ɔ ə ɛ œ œ̃ j l m ɥ w b ŋ ɲ n v g ʁ d z ʒ
	Central	ə ɐ
	Silence	sil

First, we trained a HMM/GMM system using PLP features. The three-state, cross-word triphone models were trained with the HTS variant Zen et al. (2007) of the HTK toolkit on the set of 56,231 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 used 11504 tied models, 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 by 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 75.54% to 80.02%. The mapping of Tab. 1 was used to map the phonemes of the best alignment for the training of the encoders. Each encoder was trained from 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. 2 shows classification accuracies of the phonological encoders at a frame level. In Yu et al. (2012), the *Tense* class defined for English phonology features achieved the worse classification accuracy 90.6%. Our alternative *Fortis* and *Lennis* classes performed on cross-validation set better, 96.3 and 97.0, respectively.

3.2 Phonological decoder

For the phonological decoder, we employed a DNN that can learn 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.

As a target voice, we selected a French audio book "La Comédie Humaine" of Honoré de Balzac[1],

[1] https://librivox.org/scenes-de-la-vie-privee-tome-1-by-honore-de-balzac-0812

Table 2: *Classification accuracies (%) of the French phonological encoders at a frame level.*

Class	Accuracy (%)		Class	Accuracy (%)	
	train	cv		train	cv
Labial	96.4	96.5	Nasal	98.5	98.5
Dorsal	95.1	95.2	Stop	96.4	96.6
Coronal	93.4	93.5	Approximant	96.7	96.9
Alveolar	98.1	98.1	Anterior	93.0	93.0
Postalveolar	99.4	99.4	Back	96.3	96.3
High	95.0	95.1	Lennis	97.0	97.0
Low	95.6	95.7	Fortis	96.2	96.3
Mid	95.1	95.1	Round	95.7	95.8
Uvular	97.5	97.5	Unround	94.1	94.2
Velar	98.4	98.5	Voiced	93.2	93.3
Vowel	91.7	91.8	Central	97.8	97.9
Fricative	95.4	95.4	Silence	96.9	97.1

around 21 hours long. Recordings were organised into 57 sections, and we used the sections 1–50 as a training set, 51–55 as a development (cross-validation) set and 56-57 as a testing set. The development and testing sets were 2.1 hours and 29 minutes long, respectively. The audio book was chunked into 10s long speech segments for further training and testing.

3.2.1 Training

The speech signals sampled at 16 kHz, framed by 25-ms windows with 16-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[2]. Altogether, we used the speech parametrization of 29th order as DNN output features.

Phonological posteriors z_1^k were used as DNN input features. They were extracted with the same 16-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 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.

3.2.2 Speech re-synthesis

During the phonological vocoding, the synthesized speech parameters are obtained using a forward pass on the DNN. We did not use dynamic features as smoothing of synthesized speech parameters using the pre-computed (global) variances Tokuda et al. (1995) over-smoothed the formant frequencies.

Finally, 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:

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

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

where $h(.)$ denotes nonlinear activation function (forward propagation) of the trained phonological decoder, and $z_1^k(n)$ are phonological posteriors for time n.

3.3 Pruning of the phonology posteriors

Phonological features are composed from primes that categorise a particular phoneme. Each phoneme is represented by a small number of primes, and this property can be used to compress the features with a high compression ratio. The compression of transmitted parameters was achieved by (i) setting a threshold α for pruning the parameters, and (ii) scalar quantization of the parameters. The pruning was performed using the following rule:

$$z_k = \begin{cases} 0 & \text{if } z_k \leqslant \alpha \\ z_k & \text{if } z_k > \alpha \end{cases} \tag{2}$$

and a sequence of transmitted parameters was obtained by quantizing z_k into discrete values:

$$\hat{z}_k = Q(z_k), \quad \hat{z}_k = \{0, 1, \ldots, M-1\}, \tag{3}$$

where $Q(.)$ denotes an operation of scalar quantization and M is the number of quantization levels, linearly spaced from α to 1.0.

3.3.1 Results

We evaluated the vocoder with α in the range of $[0.05, 0.35]$ with a step of 0.05. For scalar quantization we used the qantization with $M = 2^q$ levels ranged of $[1, 256]$, equivalently with the quantization bits q ranged of $[1, 8]$. The evaluation did not include F0 transmission, as we found that the DNN did not model the pitch stream adequately. It may be due to a (sub)phonetic nature of the phonological features, while F0 modelling requires supra-segmental features as well. Therefore we used in further evaluation the original F0.

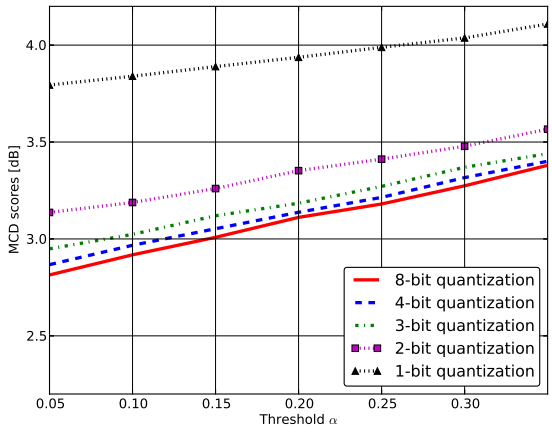


Figure 2: *Impact of phonological posteriors pruning and quantization on speech coding degradation.*

Mel Cepstral Distortion (MCD) Kubichek (1993) on the test set was used as an objective metric for evaluating the impact of pruning and quantization of the phonological posteriors on quality of speech coding. Figure 2 shows MCD scores on mel-cepstral vectors of unpruned and pruned/quantized encoded speech for different α and q. The degradation is almost linearly dependant on the increasing threshold α, with relative degradation around 0.3–0.6 dB. The quantization, except the binary case $q = 1$, where a phonology posteriors take only binary values 0/1 (so all posterior probabilities were rounded to 0/1 in respect to α), has smaller impact, around 0.15–0.3 dB.

We estimated also the transmission rates in bps for the evaluated cases. The increasing pruning from $\alpha = 0.05$ to 0.35 resulted in decreasing average number of transmitted parameters. For example, the transmission rate for $\alpha = 0.3$ and $q = 2$ was 3.7x2x62.5x5=2.3 kbps, where 3.7 is the average numbers of parameters, 62.5 is the number of frames per second, and 5 is the number of bits required to transmit

the indices of the parameters. Fig. 3 shows transmission rates for different pruning and quantization schemes. The compression ratio of transmitted parameters is almost 10:1 between 8-bit and 1-bit scalar quantizations, without audible changes in speech quality.

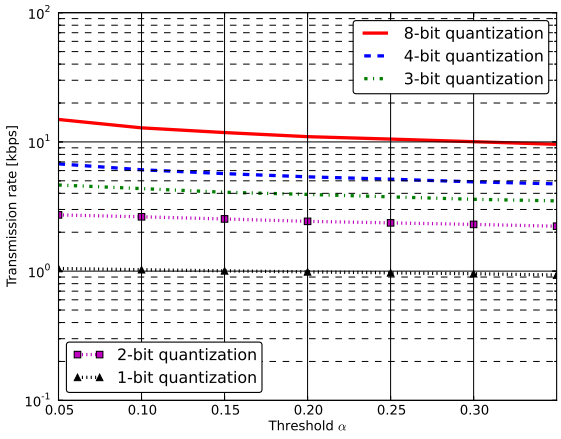


Figure 3: *Estimated transmission rate of different quantization schemes.*

We informally subjectively evaluated the degradation of speech quality for different pruning schemas and as we did not perceive differences, we did not follow with listening tests. Tab. 3 demonstrates recordings of the vocoder performance and an impact of pruning and quantization on speech quality. It is interesting that the speech quality degradation comparing to the original recording is audible but similar even for the highest compression with $q = 1$. We can conclude that the phonology posteriors form a robust speech representation, and it is quite tolerant to exact values. We speculate that it might be due to a binary nature of the distinctive features, where it is more important if a critical feature is present, and less dependent on its value.

Table 3: *Recordings demonstrating vocoder performance of encoded French testing sentence, using different pruning and quantization schemes.*

Pruning/quantization schemes	Rate [kbps]	Example
Original	–	(mp3)
$\alpha = 0.3$, $q = 8$	10	(mp3)
$\alpha = 0.3$, $q = 4$	4.9	(mp3)
$\alpha = 0.3$, $q = 2$	2.3	(mp3)
$\alpha = 0.3$, $q = 1$	0.9	(mp3)

4 Discussion and conclusions

We have shown that DNN based ASR and TTS can be cascaded to form a vocoder. Instead of full ASR system we used only phonological posteriors calculation, without a lexicon, a language model and the search module. The DNN with the open source LPC re-synthesis replaced the incremental HTS synthesis Astrinaki et al. (2012) with licensed STRAIGHT re-synthesis Kawahara et al. (1999) proposed by our previous work. This work differs from previous attempts by (i) replacing HMMs by DNNs, and (ii) using the phonological speech representation instead of the phonetic one. In the case of the real-time incremental vocoding, HMMs impose phonetic sequence constraints that leads to decrease of bit rates, but according to our experiments with the incremental vocoding, they simultaneously generate weaker phoneme boundaries that increases speech quality degradation. HMMs-based approach also requires building of a HTS voice that is not more needed in the DNN based vocoding, that is in principle unsupervised. Further, although the present validation has only been done in French, the approach is fundamentally multi-lingual. This is in contrast to the phonetic vocoder that requires a-priori knowledge of phone sets. Although the resulting bit rate of the phonological vocoder is not as low as an equivalent

phonetic vocoder, the speech quality is the same or better. However, the work is preliminary, and suggests that further research could lead to comparable bit rates. We used naive approach to encode indices of parameters, each with 5 bits, and we did not investigate other compression techniques (e.g., how often consecutive frames are identical), as the paper focused on the vocoder. In addition, using more compact phonological systems such as Chomsky's or Pöchtrager's Pöchtrager (2006) ones, only 4 or 3 bits respectively, would be required.

The work on phonological decoder is related to the DNN training for parametric TTS, described e.g. by Qian et al. (2014). In our work, we used phonological features inferred from speech signal instead of rich contextual features inferred from text. Phonology posteriors form a robust speech representation, highly tolerant to exact values. We showed that even with the highest quantization ratio using $q = 1$, i.e., using only binary values for the posteriors, the speech quality is maintained. That implies that the phonological decoder could also form a basis of a phonological parametric TTS synthesis to generate speech from the canonical phone representation. A new voice can be built quite easily and without transcripts.

Presented speech vocoding is based on high quality reconstruction of line spectra and glottal signal parameters from phonological representation. A recent study Raitio et al. (2014) showed a DNN modelling of a source signal from 47 acoustical features, while we have successfully modelled the source signal from 24 pseudo-phonological features. It is interesting to further study the impact of the critical phonological features on the source signal.

An integrated phonology encoder and decoder with transmission parameter pruning can be used as a low bit rate parametric speech coder based on a phonology representation. Related to the recent work of Flanagan Flanagan (2010), who proposes a parametric speech coding based on articulatory representation, the proposed phonological vocoder can be easily exploited if an articulatory phonology Browman and Goldstein (1986) is used as an alternative speech representation. However, it is beyond the present study. The approach of Flanagan (2010) is computationally very expensive (around 100 times real-time only to compute solutions of the Navier-Stokes fluid flow equations). The proposed phonological vocoder can operate on 1–2 kbps as well, and it is in addition real-time.

5 Acknowledgements

This research was partly supported under the RECOD project by armasuisse, the Procurement and Technology Center of the Swiss Federal Department of Defence, Civil Protection and Sport.

References

M. Astrinaki, N. d'Alessandro, B. Picart, T. Drugman, and T. Dutoit. Reactive and continuous control of HMM-based speech synthesis. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 252–257. IEEE, December 2012. ISBN 978-1-4673-5125-6. doi: 10.1109/slt.2012.6424231. URL http://dx.doi.org/10.1109/slt.2012.6424231.

Jacqueline Bauman-Waengler. *Articulatory and Phonological Impairments: A Clinical Focus (4th Edition) (Allyn & Bacon Communication Sciences and Disorders)*. Pearson, 4 edition, March 2011. ISBN 0132563568. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0132563568.

Catherine P. Browman and Louis M. Goldstein. Towards an articulatory phonology. *Phonology*, 3:219–252, May 1986. ISSN 1469-8188. doi: 10.1017/s0952675700000658. URL http://dx.doi.org/10.1017/s0952675700000658.

M. Cernak, P. Motlicek, and P. N. Garner. On the (UN)importance of the contextual factors in HMM-based speech synthesis and coding. In *Proc. of ICASSP*, pages 8140–8143. IEEE, May 2013a. doi: 10.1109/icassp.2013.6639251. URL http://dx.doi.org/10.1109/icassp.2013.6639251.

Milos Cernak, Xingyu Na, and Philip N. Garner. Syllable-Based Pitch Encoding for Low Bit Rate Speech Coding with Recognition/Synthesis Architecture. In *Proc. of Interspeech*, pages 3449–3452, August 2013b. URL http://www.isca-speech.org/archive/interspeech_2013/i13_3449.html.

Milos Cernak, Alexandros Lazaridis, Philip N. Garner, and Petr Motlicek. Stress and Accent Transmission In HMM-Based Syllable-Context Very Low Bit Rate Speech Coding. In *Proc. of Interspeech*, pages 2799–2803, September 2014.

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

J.L. Flanagan. Parametric representation of speech signals [dsp history]. *IEEE Signal Processing Magazine*, 27(3):141–145, 2010. ISSN 1053-5888. doi: 10.1109/MSP.2010.936028.

S. Galliano, E. Geoffrois, G. Gravier, J. f. Bonastre, D. Mostefa, and K. Choukri. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006*, pages 315–320, 2006.

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.

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

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.

H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Proc. of Eurospeech*, Budapest, Hungary, 1999.

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.

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.

Peter Ladefoged and Keith Johnson. *A Course in Phonetics*. Cengage Learning, 7 edition, January 2014. ISBN 1285463404. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1285463404.

Heng Lu, Simon King, and Oliver Watts. Combining a Vector Space Representation of Linguistic Context with a Deep Neural Network for Text-To-Speech Synthesis. In *Proc. of 8th ISCA Workshop on Speech Synthesis*, pages 281–285, 2013.

G. Perennou. B.D.L.E.X. : A data and cognition base of spoken French. In *Proc. of ICASSP*, volume 11, pages 325–328, 1986.

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.

M. A. Pöchtrager. *The Structure of Length*. PhD thesis, Universität Wien, Vienna, 2006.

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.

Yao Qian, Yuchen Fan, Wenping Hu, and F. K. Soong. On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis. In *Proc. of ICASSP*, pages 3829–3833. IEEE, May 2014. doi: 10.1109/icassp.2014.6854318. URL http://dx.doi.org/10.1109/icassp.2014.6854318.

Tuomo Raitio, Heng Lu, John Kane, Antti Suni, Martti Vainio, Simon King, and Paavo Alku. Voice source modelling using deep neural networks for statistical parametric speech synthesis. In *Proc. of EUSIPCO*, Lisbon, Portugal, September 2014.

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.

Xiang Yin, Zhen-Hua Ling, and Li-Rong Dai. Spectral modeling using neural autoregressive distribution estimators for statistical parametric speech synthesis. In *Proc. of ICASSP*, pages 3824–3828. IEEE, May 2014. doi: 10.1109/icassp.2014.6854317. URL http://dx.doi.org/10.1109/icassp.2014.6854317.

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.

Heiga Ze, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In *Proc. of ICASSP*, pages 7962–7966. IEEE, May 2013. doi: 10.1109/icassp.2013.6639215. URL http://dx.doi.org/10.1109/icassp.2013.6639215.

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.