



On Compressibility of Neural Network Phonological Features for Low Bit Rate Speech Coding

Afsaneh Asaei¹, Milos Cernak¹, Hervé Bourlard^{1,2}

¹Idiap Research Institute, Martigny, Switzerland

²Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

{afsaneh.asaei, milos.cernak, herve.bourlard}@idiap.ch

Abstract

Phonological features extracted by neural network have shown interesting potential for low bit rate speech vocoding. The time span of phonological features is wider than that of the phonetic features, and thus fewer frames need to be transmitted. Moreover, the binary nature of phonological features enables a higher compression ratio at minor quality cost.

In this paper, we study the compressibility and structured sparsity of the phonological features. We propose a compressive sampling framework for speech coding and sparse reconstruction for decoding prior to synthesis. Compressive sampling is found to be a principled way for compression in contrast to the conventional pruning approach; it leads to 50% reduction in the bit-rate for better or equal quality of the decoded speech. Furthermore, exploiting the structured sparsity and binary characteristic of these features have shown to enable very low bit-rate coding at 700 bps with negligible quality loss; this coding scheme imposes no latency. If we consider a latency of 256 ms for supra-segmental structures, the rate of 250 – 350 bps is achieved.

Index Terms: Very low bit rate speech coding, Phonological features, Compressive sampling, Structured sparsity, Binary representation.

1. Introduction

Current conventional low bit rate speech coders operate on 1–2 kpbs (bits per second) bit rate, achieving an annoying speech degradation. To achieve lower bit rates, parametric speech coders were proposed, and cascaded phone-based automatic speech recognition (ASR) and text-to-speech (TTS), as described, e.g., by [1], [2] and [3], became popular. We have recently contributed to the very low bit rate coding by our proposal of syllable-context phonetic decoding [4]. The system operates on 200–300 bps incrementally with a syllable latency.

Recently we have proposed to use phonological vocoder instead of phonetic one [5], based on modelling of the abstract segmental phonological features, such as defined by [6] or [7]. The motivation was to benefit from wider temporal span of the phonological features, and the underlying phonology mechanisms that define the phonological features as binary. For example, a sound [m] is realised as a binary combination of [+anterior], [+voice] and [+nasal] phonological features. Considering multilingual parametric speech coding, the use of these features seems to be also promising. The recognition/synthesis system operates on 1–3 kpbs, and it is based solely on neural networks, rather than on hidden Markov models.

The research leading to this result is supported in part by SNSF project on “Parsimonious Hierarchical Automatic Speech Recognition (PHASER)” grant agreement number 200021-153507.

The quantisation and compression of the phonological features are the main factors that determine the operating bit rate and speech quality. In [5], we proposed to prune the phonological features smaller than a certain (empirically tuned) threshold to attain higher compression. Although this pruning scheme seems to be effective, it is not suitable for codec implementation as it introduces bursts of features and highly variable code length that could impact the latency of speech coding. This paper focuses on compression of the phonological features. Relying on sparsity of these features, we propose to apply a *compressive sampling* method to provide a low-dimensional projection of these features. This approach leads to fixed length codes for transmission so it is very convenient for codec implementation. At the decoding state, the sparse phonological features are reconstructed using *sparse recovery* algorithm. The compressive sampling and sparse reconstruction scheme define a principled way for phonological vocoding. The experimental analysis demonstrates up to 50% bit-rate reduction compared to the conventional pruning approach for better or equal quality of the decoded speech.

Furthermore, we study the structured sparsity of the phonological features. The intuition is that the phonological features lie on low-dimensional subspaces. The low-dimension pertain to either *physiology* of the speech production mechanism or the *semantic* of the supra-segmental information. At the physiology level, only certain (very few) combinations of the phonological features can be realized through human vocalization. This property can be formalized by constructing a codebook of structured sparse codes for phonological feature representation. Likewise, at the semantic level, only certain (very few) supra-segmental (e.g. syllabic) mapping of the sequence of phonological features is linguistically permissible. This property can be exploited for block-wise coding of these features with a slower (supra-segmental) dynamic. We demonstrate that structured sparse coding of the binary features enables the codec to operate at 700 bps without imposing any latency or quality loss with respect to the earlier developed vocoder [5]. By considering a latency of about 256 ms, the bit rate of 250 – 350 bps is achieved without requirement for any prior knowledge on supra-segmental (e.g. syllabic) identities.

The rest of this paper is organized as follows: some background information on compressive sampling (CS) is briefly stated in Section 2. In Section 3, we analyze the compressibility of the phonological features from the CS perspective. The CS-based vocoder is described in Sections 4. We study the structured sparsity of the phonological features and its implications for designing an efficient vocoder in Section 5. The experimental results are presented in Section 6 and the conclusions are drawn in Section 7.

10.21437/Interspeech.2015-167

2. Compressive Sampling Principles

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

$$z = D \alpha \quad (\text{CS coder}) \quad (1)$$

suffice to reconstruct α , where $D \in \mathbb{R}^{M \times N}$ is a *compressive measurement matrix* which preserves the pairwise distances of the sparse features α in the compressed code z . Given an observation vector z , and the measurement matrix D , the sparse representation α is obtained by the optimization problem stated as

$$\min_{\alpha} \|\alpha\|_0 \quad \text{subject to} \quad z = D \alpha \quad (\text{Sparse decoder}) \quad (2)$$

where the counting function $\|\cdot\|_0 : \mathbb{R}^M \rightarrow \mathbb{N}$ returns the number of non-zero components in its argument. The non-convex objective $\|\alpha\|_0$ is often relaxed to $\|\alpha\|_1 = \sum_i |\alpha_i|$ which can be solved in polynomial time [8]. Recent advances in CS exploits inter-dependency structure underlying the support of the sparse coefficients in recovery algorithms to reduce the number of required observations and to better differentiate the true coefficients from recovery artifacts for higher quality [9].

3. Compressibility of Phonological Features

To investigate how the Phonological features fit a common metric for sparsity, we analyze the power-law decay of these representations [10]. For the features to be closely approximated as sparse thus compressible, the coefficients α must have a rapid power-law decay when sorted:

$$|\alpha_i| \leq \gamma i^{-\frac{1}{r}} \quad r \leq 1, \quad (3)$$

where $\alpha_i, \forall i \in \{0, \dots, N\}$ denotes the coefficients of α when sorted from largest to smallest. Plotting the sorted value of the phonological features vs. their index is illustrated in Figure 1.

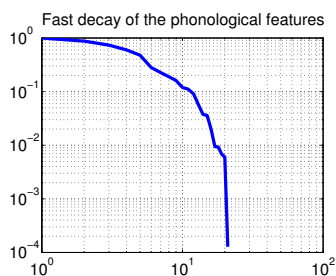


Figure 1: Compressibility of the phonological features conforming to the power-law decay.

We can see that the decay of the coefficients satisfies equation (3). Based on this observation, the phonological representation of speech signal is considered to be *compressible* and the compressive sampling can be effectively used to reduce the dimension of the features for coding while sparse recovery guarantees reconstruction of the original sparse coefficients at the decoding. In this study, $N = 25$. In addition, the PCA analysis of the phonological features reveals that only the first 12 coefficients (out of 25) capture above 95% of the variability.

4. CS-based Phonological Vocoding

Relying on the compressibility of the phonological representation of speech signal, we propose to apply compressive sampling to reduce the dimension of these features. Earlier work [5] used pruning based on a threshold value which has to be tuned for different conditions. This approach is not practical as the resulting code has a variable length and leads to burst of features and latency for codec implementation. In contrast, compressive sampling lays out a principled way for dimensionality reduction of sparse representation and it is a practical framework for coding. Figure 2 shows the functional blocks of the proposed speech coder.

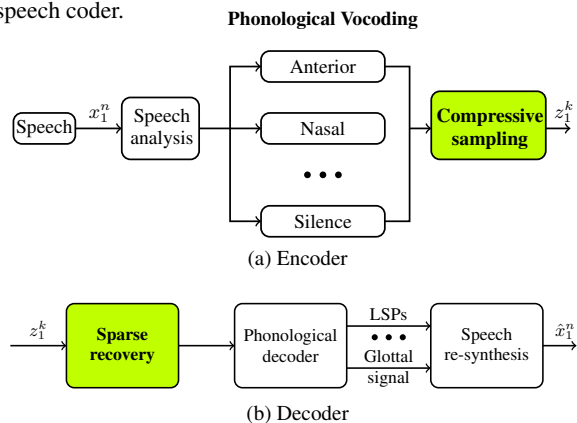


Figure 2: CS-based phonological vocoder split into (a) encoder and (b) decoder. Compressed features are recovered at the receiver side where the decoder generates speech spectra lines LSPs and source parameters for speech re-synthesis.

The proposed CS-based phonological vocoding consists in two steps: (i) CS-coder as expressed in (1) and (ii) Sparse recovery for decoding as presented in its general form in (2).

4.1. CS Coder

At the coding step, the choice of *compressive measurement matrix* D is very important. A sufficient but not necessary condition on D to guarantee decoding of the sparse representation coefficients is that all pairwise distances between K -sparse representations must be well preserved in the observation space or equivalently all subsets of K columns taken from the measurement matrix are nearly orthogonal. This condition on the compressive measurement matrix is referred to as the restricted isometry property (RIP). The random matrices generated by sampling from Gaussian or Bernoulli distributions are proved to satisfy RIP condition [11]. To generate D in the Gaussian case, we generate samples from a multivariate Gaussian distribution. On the other hand, we can create a binary matrix D by setting around 50% of the components of each column at random permutations to 1 [12]. We study the empirical difference between the two compression mechanisms, i.e. Gaussian vs. Bernoulli in Section 6 where a choice of Bernoulli matrix is demonstrated to achieve higher robustness to quantization.

Given the compressed codes, there are infinitely many solutions to reconstruct the original high-dimensional representation which satisfy (1). Relying on the two principles of (1) sparse representation and (2) incoherent measurement, we can guarantee to circumvent the ill-posedness of the problem and recover the K -sparse data stably from the compressed (low-dimensional) observations through efficient optimization algorithms which search for the sparsest representation that agrees with those observations. This step is implemented at the decoder [11].

4.2. Sparse Decoder

At the decoder, the high-dimensional phonological features are reconstructed using constrained LASSO sparse recovery algorithm [8] expressed as

$$\begin{aligned} \hat{\alpha} = \arg \min & \|\alpha\|_1 + \lambda \|z - D\alpha\|_2 \\ \text{subject to} & \quad 0 < \alpha < 1 \end{aligned} \quad (4)$$

where λ is the regularization parameters. The first term $\|\cdot\|_1$ is a relaxed (convex) version of the ℓ_0 semi-norm sparse recovery problem stated in (2). This term promotes the sparsity of the recovered representation. This term can be replaced by $\|\cdot\|_\infty$ standing for the ℓ_∞ -norm defined as the maximum component of α . It is shown that ℓ_∞ -norm leads to de-quantization effect [13].

The second term in (4) accounts for the reconstruction error. Regularization on the ℓ_2 -norm is equivalent to the solving the constrained optimization, $z = D\alpha$ if the measurements are not quantized. The constraint $0 < \alpha < 1$ is set for the phonological features as they are neural network estimated posterior probabilities for each individual phonological class. Having the prior knowledge of the bound of the features, eliminate the need for ℓ_∞ -norm as verified empirically in the experimental analysis presented in Section 6.

5. Structured Sparse Binary Coding

The phonological features are indicators of the physiological posture of the human articulation machinery. Due to the physical constraints, only few combinations can be realized in our vocalization. This physical limitation leads to a small number of unique patterns exhibited over the entire speech corpora. We refer to this structure as *physiological structure* which is exhibited at a frame level.

In addition, there is a block (repeated) structure underlying a sequence of phonological features. This structure is exhibited at the supra-segmental level by analyzing a long duration of the features. This structure is associated to the syllabic information underlying a sequence of phonological features. We refer to this structure as *semantic structure*. Figure 3 illustrates the structured sparsity and binary characteristic of the phonological features.

The structured sparsity of the phonological features enables us to construct a codebook for very efficient coding. To this end, we consider *binary* phonological features that have been shown efficient for very low bit rate speech coding [5]. As a case study, we use the features generated for an audiobook with the length of 21 hours speech. The total number of unique structures emerging out of total number of 4746186 frames is only 12483 which is about 0.26% of the whole features. By identifying all the unique *binary* structures, a codebook is constructed for phonological feature representation. It is evident that only 14 bits are enough for transmitting a code. Given that the number of frames per second for phonological vocoding is 50^1 , this coding scheme leads to $50 \times 14 = 700$ bits per second transmission rate.

Furthermore, from a supra-segmental view, there is strong correlation between the adjacent features due to limited permissible linguistic combinations. The supra-segmental linguistic

¹It may be noted that in the phonological vocoding system, a neural network is trained for silence detection [5], thus the silence/pause intervals are coded efficiently with a small overhead. Because around 20% of the transmitted speech is detected as silence (valid for our training data), we obtain the effective speech frame-rate as $62.5 \times 0.2 = 50$.

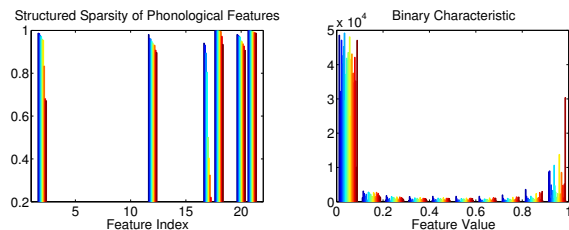


Figure 3: **Left:** Phonological features exhibit structured sparsity at two distinguished levels: (1) Physiological structure: The histogram shows activation of phonological features at (2,12,17,18,20,21); this pattern encodes a particular shape of the vocal tract. Since a limited number of these shapes can be created for human speech, the number of unique patterns is very small. (2) Semantic structure: The histogram shows that activations of the phonological pattern (2,12,17,18,20,21) is persistent through time as it gets repeated in consecutive frames. This pattern encodes the duration of each physiological shape leading to a supra-segmental capturing of syllabic information. This property encourages structured sparse coding of these features. **Right:** Phonological features have a binary nature in which most of components have either very large values close to one or very small values close to zero. This property encourages binary approximation of these features.

units may correspond to the syllabic identities or stressed regions. While exploiting the supra-segmental information has been shown to yield significant bit-rate reduction [4], in practice, providing the syllabic information requires additional processing which can impose higher cost on the codec. On the other hand, constructing a codebook of structured sparse binary patterns as described above is straightforward and requires less analysis. The supra-segmental information can be captured by imposing a latency and transmitting the blocks repeated patterns. As a case study, investigating the features obtained for the audiobook reveals that the number of blocks is less than 36% of the total number of frames and 4 bits is sufficient to transmit the number of repeated codes. That amounts to $0.36 \times 50 \times (14 + 4) = 328$ bps transmission rate with no loss in the quality of the reconstructed speech. If the duration information is dropped, then the bit rate is only 250 bps; further analysis is required to evaluate the extent of distortion that ignoring the temporal duration can impose on ineligibility of the reconstructed speech.

6. Experimental Analysis

The experiments are conducted to study the performance of the proposed phonological vocoding systems at different bit-rates in terms of objective and subjective quality measures. Different bit-rates are achieved through linear quantisation of the transmitted codes.

6.1. Phonological Encoder and Decoder Setup

The experimental setup used in this work follows the setup used in [5]. Briefly, at the encoder, a bank of phonological classifiers is realised using neural network to generate the posterior probability of the input acoustic vector belong to the individual phonological classes. Overall $N = 25$ phonology classes are considered and 25 separate neural networks are trained to generate the binary phonological posterior for each individual class. The French speech database Ester [14] of standard French radio broadcast news was used for training of the encoders. The architecture of the neural network was determined empirically that led to 3-hidden layer of dimensions 2000x500x2000, trained on the 16 kHz speech signals, framed by 25-ms windows with 16-

ms frame shift, using temporal context of 9 successive frames of PLP features, and softmax output function.

At the decoder, a DNN is used to learn the highly-complex regression problem of mapping phonological features 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 ², 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 set and 56 – 57 as a testing set. The development and testing sets were 2.1 hours and 29 minutes long, respectively. The DNN was initialised using 4x1024 deep belief network (DBN) pre-training by contrastive divergence with 1 sampling step (CD1) [15]. 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 [16]. Finally, speech was re-synthesised using an *open-source* LPC vocoder based on minimum-phase complex cepstrum glottal model estimation [17]. 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. If we consider pitch transmission and a latency of 256ms that corresponds to an average syllable duration, we can use the syllable-based pitch coding [18], operating on 30–40 bps.

6.2. Reconstructed Speech Quality

To evaluate the reconstructed speech quality, signal to noise ratio (SNR) and Mel cepstral distortion (MCD) [19] are used as objective metrics. In addition, the overall quality of the proposed speech coding is evaluated subjectively using the degradation category rating (DCR) procedure [20] quantifying the degradation mean opinion score (DMOS). This method provides a quality scale of high resolution, due to comparison of a distorted (synthesized) signal with a (natural/original) reference. The test consisted of 8 sentences randomly chosen from the 57th (testing) section of the audio book, with length of at least 2 seconds. Twelve listeners were asked to rate the degradation of encoded speech samples compared with reference signals based on their overall perception. According to the DCR procedure, it is not fair to build a pair associating two encoded samples since it would have implied that the first encoded sample outclasses perception of the second one. Therefore natural speech was selected as a reference sample in the test. Listeners had to describe degradation within the following five categories: [1]: Very annoying, [2]: Annoying, [3]: Slightly annoying, [4]: Audible but not annoying, [5]: Inaudible.

The evaluated synthesis systems operate at different bit rates by applying the linear quantisation at 8-level (Q=8), 4-level (Q=4), 3-level (Q=3) using the compressive sampling and sparse reconstruction scheme (Section 4). In addition, the structured sparse binary coding scheme as described in Section 5 is evaluated which operates at two bit rates for transmission of binary features at *no latency* (Q=1) and *256 ms imposed latency* (Q'=1). According to the G.114, the users are “satisfied” as long as latency does not exceed 280 ms [21]. Table 1 lists all the results. A *t*-test confirmed that the differences between the coding schemes are statistically significant ($p < 0.05$). We can see that the phonological vocoder can achieve very low bit rate at a minor quality loss thanks to their binary charac-

Table 1: Quality evaluation results for reconstructed speech for different transmission bit rates obtained for various linear quantisation regimes; Q indicates the number of quantization bits. All the systems impose no latency except for (Q'=1) that requires a latency of 256 ms.

Quantz	SNR	MCD	DMOS	Bit rate [bps]
Q'=1	10.4	3.8	2.20	328
Q=1	10.4	3.8	2.20	700
Q=3	9.60	4.04	2.08	2300
Q=4	13.3	3.68	2.35	3050
Q=8	20.2	2.87	2.84	6050

teristic and strong physiological and semantic structure underlying their sparse representation. The DMOS scores compare with quality of the state-of-the-art low bit rate speech coders. Furthermore, the compressive sampling scheme leads to 50% bit-rate reduction compared to the pruning approach employed in [5]. Although at the fine quantisation regime (e.g. Q=8), the objective and subjective quality of the reconstructed speech is far better than the pruning method [5], compressive sampling and sparse reconstruction is found sensitive to quantization of the measurements so at the $Q = 3$ quantization level, the quality of both approaches are indistinguishable while CS leads to higher compression. Hence, compressive sensing and sparse recovery demonstrate high sensitivity to quantization. Although the use of ℓ_∞ -norm has been shown to yield some level of de-quantization, it did not have any significant impact on our phonological vocoding system. It can be justified due the binary nature of these features and bounded constrained optimization defined in (4). On the other hand, the choice of compressive sensing matrix is crucial for robustness against quantization. Table 2 shows the quality of the reconstructed speech in terms of SNR (dB) for the choice of Bernoulli contrasted with Gaussian random matrices.

Table 2: Impact of compressive measurement matrix on reconstructed speech quality quantified in terms of SNR (dB).

Quantization	Bernoulli	Gaussian
Q=8	20.2	17.6
Q=4	13.3	9.5
Q=3	9.60	5.37

We can see that the Bernoulli compressive measurements lead to smaller degradation. The differences are more pronounced at highly quantized regimes such as $Q = 3$.

7. Conclusions

Compressive sensing and sparse recovery are found to be the systematic way of compression and recovery of the sparse phonological features. Moreover, structured sparsity along with binary approximation of the phonological features are exploited for construction of a phonological codebook that encapsulates the postures of human vocalisation. This coding scheme operates at 700 bps with very small subjective degradation in quality of the reconstructed speech compared to unquantized features. Considering the supra-segmental information by imposing some latency at the codec, very low-bit rate of 324 bps is achieved. While the compressibility of the phonological features is impressive, utilizing a DNN at the decoder and the LPC vocoder have a great impact on the quality of the reconstructed speech. Adaption of the DNN along with a better vocoder can diminish some effect of reconstruction and yield higher speech quality.

² librivox.org/scenes-de-la-vie-privee-tome-1-by-honore-de-balzac-0812

8. References

- [1] J. Picone and G. R. Doddington, "A phonetic vocoder," in *Proc. of ICASSP*. IEEE, May 1989, pp. 580–583 vol.1. [Online]. Available: <http://dx.doi.org/10.1109/icassp.1989.266493>
- [2] 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. [Online]. Available: <http://dx.doi.org/10.1109/icassp.1998.675338>
- [3] 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.
- [4] 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, 2015.
- [5] M. Cernak, B. Potard, and P. N. Garner, "Phonological Vocoding Using Artificial Neural Networks," in *Proc. of ICASSP*, Apr. 2015, pp. 4844–4848. [Online]. Available: <http://publications.idiap.ch/index.php/publications/show/3070>
- [6] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, NY: Harper & Row, 1968.
- [7] J. Harris, *English Sound Structure*, 1st ed. Wiley-Blackwell, Dec. 1994. [Online]. Available: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0631187413>
- [8] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab software for disciplined convex programming," 2008.
- [9] A. Asaei, H. Bourlard, and V. Cevher, "Model-based compressive sensing for multi-party distant speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.
- [10] V. Cevher, P. Indyk, L. Carin, and R. G. Baraniuk, "Sparse signal recovery and acquisition with graphical models," *Signal Processing Magazine, IEEE*, vol. 27, no. 6, pp. 92–103, 2010.
- [11] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21–30, 2008.
- [12] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," in *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, 2008, pp. 798–805.
- [13] L. Jacques, D. K. Hammond, and M.-J. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 559–571, 2011.
- [14] 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)*, 2006, pp. 315–320.
- [15] 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. [Online]. Available: <http://dx.doi.org/10.1162/neco.2006.18.7.1527>
- [16] 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.
- [17] P. N. Garner, M. Cernak, and B. Potard, "A simple continuous excitation model for parametric vocoding," Idiap, Tech. Rep. Idiap-RR-03-2015, Jan. 2015. [Online]. Available: <http://publications.idiap.ch/index.php/publications/show/2955>
- [18] 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. [Online]. Available: http://www.isca-speech.org/archive/interspeech/_2013/i13/_3449.html
- [19] 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. [Online]. Available: <http://dx.doi.org/10.1109/pacrim.1993.407206>
- [20] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," (Geneva, Switzerland) 1996.
- [21] ITU-T Rec. G.114, "One-way transmission time ," (Geneva, Switzerland) 2003.