Quantized Posterior Hashing: Efficient Posterior Exemplar Search Exploiting Class-Specific Sparsity Structures

Afsaneh Asaei[1], *Gil Luyet*[1,3], *Milos Cernak*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3]University of Fribourg, Switzerland
{aasaei,gluyet,mcernak,bourlard}@idiap.ch

Abstract

This paper shows that exemplar-based speech processing using class-conditional posterior probabilities admits a highly effective search strategy relying on posteriors' intrinsic sparsity structures. The posterior probabilities are estimated for phonetic and phonological classes using deep neural network (DNN) computational framework. Exploiting the class-specific sparsity leads to a simple quantized posterior hashing procedure to reduce the search space of posterior exemplars. To that end, small subset of quantized posteriors are regarded as representatives of the posterior space and used as hash keys to index buckets of similar exemplars. The k nearest neighbor (kNN) method is applied for posterior based classification problems. The phonetic posterior probabilities are used as exemplars for phoneme classification whereas the phonological posteriors are used as exemplars for higher level linguistic parsing. Experimental results demonstrate that posterior hashing improves the efficiency of kNN classification drastically. This work encourages the use of posteriors as discriminative exemplars appropriate for large scale speech classification tasks.

Index Terms: Fast kNN, Structured sparsity, Quantized posterior hashing, Posterior representatives, Phoneme classification, Linguistic parsing.

1. Introduction

Exemplar-based speech processing provides a powerful big data solution for potentially a wide range of speech technologies. In particular, speech classification relying on exemplar matching possess higher flexibility than the statistical methods due to lack of prejudices on data and expected answers. The fundamental question that yet remains is the application-specific appropriate choice of exemplars. The present manuscript reinforces the position of deep neural network (DNN) based class-conditional posterior probabilities (briefly referred to as *posteriors*) as a great choice of exemplars for speech classification tasks.

In theory, if infinite number of exemplars of continuous probability density functions are provided, a simple nearest-neighbor rule leads to optimal classification [1]. In the context of speech recognition, the nearest-neighbor based techniques have been used as non-parametric methods to perform class-conditional probability estimation for acoustic modeling [2, 3]. Typical choice of exemplars are variants of spectral features [4, 5], and approximate neighborhood search strategies are tailored to provide tractable frameworks [2].

Application of hashing in nearest neighbor search enables splitting the search space into buckets each identified with a unique hash key. The exhaustive search space is thus downsized to the corresponding bucket sizes [6]. In this context, the hash function ensures locality preserving of similar/neighboring data while the whole space is spanned in disjoint splits [7]. In posterior space, all above essential features are obtained through a simple *linear posterior quantization* to generate the hash keys as posterior representatives, and populate the corresponding buckets of similar posterior vectors.

DNN posteriors live in union of low-dimensional/structured sparse subspaces [8, 9]. Exploiting this property enables a hierarchical speech classification and recognition framework based on sparse modeling of phonetic posterior exemplars [10]. In addition, the low-dimensional subspaces can be modeled through dictionary learning for sparse representation where projection of the posteriors into the space characterized over the training data reduces the mismatch of the testing posteriors, and leads to enhanced acoustic modeling for speech recognition [8, 9]. Another application of DNN phonetic posterior exemplars have been established in query-by-example spoken term detection [11], and sparse subspace modeling have been found promising to address this problem [12, 13].

In addition to the phonetic posteriors, our previous studies on phonological posteriors show that they conform to a small number of unique binary structures which are a tiny fraction of the number of permissible codes. Exploiting this property enables construction of a small-size codebook for very low-bit rate speech coding [14]. More recently, we also considered structured sparsity of phonological posteriors for higher level classification of linguistic properties, also referred to as linguistic parsing [15].

In this paper, we propose a novel application of structured sparsity of posterior probabilities in devising an effective hashing technique to reduce the search space of posterior exemplars. Motivated by the idea of locality sensitive hashing for fast neighborhood search [7], the geometric locality in the space of posteriors can be defined by thresholding the high probability components at multiple quantized levels. As the variability in the space of posteriors is largely confined to the underlying class probabilities, grouping the posteriors according to their similar quantized codes ensures the actual neighbors in disjoint buckets covering the whole space. The posteriors are investigated at two levels corresponding to phonetic and phonological classes. We consider phonetic classification and phonological linguistic parsing based on k nearest neighbor (kNN) search using the quantized posterior hashing technique.

In the rest of the paper, the posterior hashing theory is described in Section 2. Experimental studies are carried out in Section 3, and the conclusions are drawn in Section 4 with an outlook to development of large scale, fast and flexible speech technologies relying on DNN posterior exemplars.

2. Quantized Posterior Hashing

Inspired from the idea of locality sensitive hashing [7], a deterministic procedure for posterior space hashing is proposed. This procedure relies on structured sparsity of posterior subspaces to characterize the geometric localities, and enables search space reduction for neighborhood analysis of posterior exemplars.

2.1. Structured Sparsity

Phonetic and phonological posterior estimation requires speech analysis that turns a sequence of N acoustic feature observations $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \dots, \boldsymbol{x}_N\}$ into a sequence of N posterior probability vectors $Z = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots, \boldsymbol{z}_N\}$ where

$$\boldsymbol{z}_n = [p(C_1|\boldsymbol{x}_n), \dots, p(C_q|\boldsymbol{x}_n), \dots, p(C_Q|\boldsymbol{x}_n)]^\top$$

consists of Q class-conditional posterior probabilities, and $.^\top$ denotes the transpose operator. DNN is the state-of-the-art computational method to estimate the posterior probabilities. A DNN learns the categorical distribution mapping an input acoustic feature $\boldsymbol{x}_n$ to a specific class probability. A single DNN is employed to estimate the phonetic posteriors, whereas multiple parallel DNNs are used for detection of different phonological classes; the DNN outputs are then concatenated to form the phonological posterior vector (details in 3.2).

Figure 1 depicts a sample of phonological and phonetic posteriors estimated for an utterance of speech signal. The left plots illustrate the binary quantized posteriors of the continuous probabilities depicted in the right plots. An exclusive class-specific sparsity is evident for phonetic posteriors which is pertained to the DNN exclusive mapping to the hard output phonetic labels. As a result, the class-specific subspaces are highly structured or the matrices of class-specific phonetic posteriors have a very low-rank [8], and the posterior space is a union of these low-dimensional subspaces.

On the other hand, the restricted multi-class probabilities are visible at phonological posteriors. As the phonological classes correspond to sub-phonetic attributes, multiple classes are activated for generation of linguistic units at any instance of time, however, their combination is confined to a small permissible activation of articulatory mechanisms as determined in a phonological system [16]. These permissible combinations define the sparsity structures underlying phonological posteriors. In both phonetic and phonological illustrations, the sparse vectors exhibit a sequencing structure inherited from the input acoustic feature observations. More details about the DNN setup for estimation of posteriors, their dimension and databases will be described in Section 3.

2.2. Quantized Posterior Representatives are Hash Keys

Taking advantage of the underlying structured sparsity of posteriors, it is possible to design a hashing technique to divide the space into smaller size buckets of locally neighbor posteriors.

As already discussed in Section 2.1, posterior exemplars are sparse vectors residing in union of low-dimensional subspaces. Hence, for any posterior vector, the probabilities are confined to a very small number of components where the indices of high probabilities identify the unique structure of the underlying subspace. Accordingly, the number of unique quantized posteriors is relatively small with respect to the sample size, and the quantized posteriors can be regarded as representatives of the posterior space. The quantized posterior representatives can be used as hash keys for splitting the space into geometric neighbors.

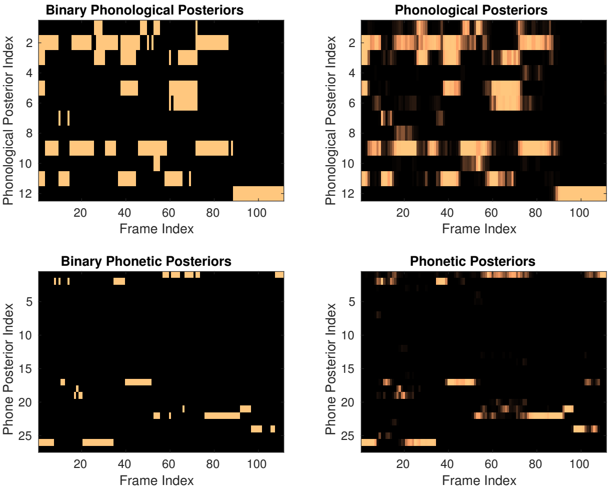

Figure 1: *Posteriograms of phonological and phonetic posteriors. Quantized posteriors are representatives of the posterior space that can be used as the hash keys for effective nearest neighbor search.*

The following hashing formula express this method through different levels of quantization.

$$H(\boldsymbol{z}) = \frac{\lfloor 2^b \boldsymbol{z} \rfloor}{2^b} \tag{1}$$

where $\boldsymbol{z}$ is a posterior probability vector, and b is the number of bits for quantization.

> *The main idea is that the quantized posterior vectors can be used directly as the hash keys to form buckets of similar posterior exemplars. The quantized posterior hashing divides the search space into disjoint buckets where the number of buckets is proportional to the number of disjoint classes.*

In theory, the number of disjoint classes is equal to the dimension of DNN outputs for phonetic posteriors, or equal to the number of permissible combinations of the phonological classes as defined in a phonological system [17, 18]. Our experimental studies conducted in Section 3 confirm this proportionality. As the number of permissible combinations is relatively large in a phonological system, we will see that even the binary posterior hashing leads to effectively small-size buckets encapsulating similar posterior exemplars. Generally, as the number of quantization bits is increased, the buckets will become more specific. By fixing a minimum bucket size, a hierarchical hashing can be devised for search space indexing, and the neighbor search is accomplished for the finest matched quantization key (details in 3.3).

This hashing technique can be combined with kNN to enable efficient posterior classification. We will see in Section 3 that this simple hashing idea can reduce the search space of posteriors tremendously with no degradation of kNN performance.

2.3. Analysis and Cost

In theory, quantization of every component of posteriors in b bits leads to splitting the space in maximum $K = 2^b Q$ disjoint regions. Accordingly, the size of training data in each bucket can be reduced to an average $N/2^K$. The minimum similarity occurs for the vectors at the boundaries, thus maximum distance of the two vectors with equal hash keys is

$$\max_{H(\boldsymbol{z_1})=H(\boldsymbol{z_2})} d(\boldsymbol{z_1}, \boldsymbol{z_2}) \simeq d(\boldsymbol{z}, \boldsymbol{z} + \frac{1}{2^b}\mathbf{1}_K)$$

where $\mathbf{1}_K$ is an all-ones vector of dimension K. Hence, the probability of negative examples in a bucket is equal to the probability of having an ℓ_1-distance smaller than $\frac{1}{2^b}$ in every dimension. Since the maximum ℓ_1 distance is 1 (posterior definition), this probability is $p = \frac{1}{2^b}$. The negative examples occur when the K-dimensional keys of the two posteriors of different classes are equivalent, therefore, negative examples have the probability of $1 - p^K$. Since the typical number of Q is often more than 40, this hashing function leads to a very small probability of encapsulating negative examples or wrong positive examples in the same bucket.

In practice, the number of non-empty buckets is very small, and by considering a large minimum number of exemplars per bucket, the number of effective hash keys for each class is proportional to $\|\lfloor 2^b E(\boldsymbol{z}|C_q)\rfloor\|_0$), where $E(.)$ is the expected value of the posteriors obtained by averaging the probability vectors belonging to the class C_q, and $\|.\|_0$ denotes the number of non-zero components. The overall number of hash keys is thus $\sum_{q=1}^{Q} \|\lfloor 2^b E(\boldsymbol{z}|C_q)\rfloor\|_0$. The experimental results are presented in Section 3.

3. Experimental Results

In this section, we evaluate the effectiveness of hashing for neighborhood search of posterior exemplars. Two kNN tasks are investigated, namely, phoneme classification using phonetic posteriors, and linguistic parsing using phonological posteriors.

3.1. DNN Setup for Phonetic Posteriors

The phoneme classification is conducted on AMI corpus [19]. A DNN is used to estimate the phonetic posterior probabilities, and it is implemented using Kaldi toolkit [20]. Its architecture consists of three hidden layers with 1024 nodes. The input features of the DNN are MFCC features plus the first and second order dynamic features using a context of 9 frames at frequency 100. The DNN outputs are hard labels corresponding to 43 dimensional English phoneme classes. The training labels are obtained from hidden Markov model (HMM) force alignment using speech transcription. We use the standard splitting of training, development and test data as performed in [21].

3.2. DNN Setup for Phonological Posteriors

The Wall Street Journal WSJ0 and WSJ1 continuous speech recognition corpora [22] are used for training the phonological class detectors. Phonological detectors are trained on the training set *si_tr_s_284* including 37,514 utterances using the Sound Pattern of English phonological features [16]. For each phonological class, a 3x1024 DNN is initialized by deep belief network pre-training of [23], and trained using Kaldi toolkit [20]. The DNN output is trained as either 1 or 0 if the phonological class is present or not. Hence, each DNN estimates the probability of occurrence of one phonological class. The outputs of all DNNs are concatenated to form a phonological posterior vector.

To perform linguistic parsing experiments, the Government Phonology (GP) posteriors [17, 18] are estimated on a labeled sub-set taken from the SIWIS database [24]. The dimension of phonological posteriors is 12 according to the GP phonological system that consists of the three basic resonance phonological primes commonly labeled as A, U, I, denoting the peripheral vowel qualities [a], [u] and [i] respectively. Other vowels are defined by a composition of the basic ones, such as [e] results in fusing the I and A primes. In addition to these 'vocalic' primes, GP proposes also the 'consonantal' primes. The evaluation data consists of recordings of 10 training and 3 testing English speakers. Each speaker reads about 25 sentences, among which 5 questions, with focus (emphasis) on one predefined word. The corresponding transcription for each sentence was given, with a tag on the words that the speakers were asked to emphasize. Hence, the goal of linguistic parsing is basically detection of emphasized segments in an utterance where the boundary of the segments are known beforehand, i.e. a top-down detection scenario.

3.3. Fast kNN for Phoneme Classification

The cosine similarity is used in kNN search for phoneme classification [25]. The value of k for exhaustive search kNN is chosen as 150; for hashing-based kNN at binary level, it is chosen as 20, and for higher levels, it is 11 using the 10-fold cross validation. Since AMI [19] is a very large corpus, we perform classification of a random selection of 50'000 test posteriors using the labeled development posteriors.

The posterior representatives or hash keys are obtained at different quantization levels. The minimum bucket size is fixed to 500, so if a hash key results in less than 500 neighborhood posteriors, it is discarded. For any test posterior, hashing can be implemented in a hierarchical procedure. The test hash keys at multiple quantization levels are obtained and compared with the available hash keys of labeled data using Jaccard distance, then the finest matched key with Jaccard distance = 0 is used to determine the bucket of neighboring posteriors. This procedure leads to multi-level hashing (mul-h). Alternatively, the single level quantization codes can be used to obtain the hash keys where the bucket with the closest key (even if the Jaccard distance is greater than 0) is used for neighborhood search. This procedure leads to single level hashing (sin-h). The results of kNN phoneme classification performance are listed in Table 1.

Table 1: *kNN Phoneme classification using single level (sin-h) and multi-level (mul-h) quantized posterior hashing on AMI database.*

#QB	#Buc.	Avg size	Acu. (sin-h)	Acu. (mul-h)
64	1	3'174'011	70.3%	70.3%
1	42	60'447	69.9%	69.9%
2	319	9'371	69.6%	69.9%

We observe that the number of unique hash keys is very small, and the search space can be reduced drastically with almost no degradation in kNN classification accuracy. In addition, the hash keys can be stored as binary vectors to enable fast binary matching to find the appropriate bucket, and parallel matching of the hash keys at multiple levels minimizes the computational overhead. The testing posteriors with different hash keys can be processed interdependently in parallel streams that can lead to higher speed up in exemplar based frameworks.

3.4. Fast kNN for Phonological Linguistic Parsing

The information of linguistic events at supra segmental level is encoded in phonological attributes, thus classification of high-level linguistic events (such as stress or syllables) is feasible exploiting the structure of high probability phonological posteriors [15]. This task is also referred to as linguistic parsing. In this section, we study emphasis detection based on nearest neighbor search using phonological posteriors along with Jaccard similarity measure.

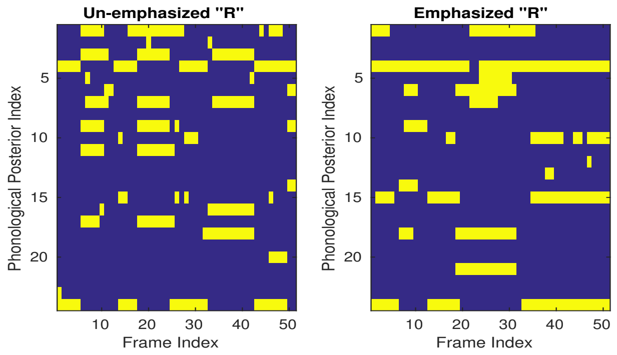

Figure 2: *Structured sparsity of phonological posteriors depicted for two pronunciation of the phoneme "R" in un-emphasized and emphasized words.*

This method relies on the hypothesis that phonological posteriors encode information about the emphasized or un-emphasized speech utterances in the support of their high probability components. The support can be identified using quantization at different levels. Figure 2 illustrates an example of binary structures underlying un-emphasized and emphasized realizations of phoneme "R". The difference in the binary patterns is evident. To visualize this property in a larger scale, we plot the t-distributed stochastic neighbor embedding (tSNE) [26] of arbitrary selection of 900 frames of phonological representations corresponding to phoneme "R" with and without emphasis in Figure 3.

This empirical observation suggests that the binary structures of phonological posteriors are indicative of their emphatic nature, thus they can be regarded as representatives of emphatic and non-emphatic variability of phonological posteriors. Accordingly, the binary structures are used as the hash keys to split the space into buckets of neighboring exemplars. The nearest neighbor rule is then used for emphasis detection (c.f. [15] for details of the implementation). The results are listed in Table 2.

Table 2: *Emphasis detection using phonological posterior hashing on SIWIS database.*

#QB	#Buckets	Avg bucket-size	Accuracy
64	1	100'284	87.1%
1	405	222	93.5%

We can see that the search space of posterior exemplars is reduced by extracting the binary codes, however, the number of buckets (unique hash keys) is more than the binary codes of phonetic posteriors. This can be explained due to the definition and training of phonological posteriors which results in high probability components corresponding to multiple phonological classes (as opposed to phonetic posteriors which often have a single high probability component).

Furthermore, unlike phonetic posterior hashing, we do not enforce any constraint on the bucket size since the best value of k is found to be 1 for this task. During the nearest neighbor search, if a binary code does not perfectly match the training hash keys, the most similar code quantified in terms of Jaccard similarity is used. We can see that *binary* hashing of phonological posteriors enables restriction of the search space to the more "correct" exemplars as the classification improves through this confined search. The results of emphasis detection are better than a state-of-the-art emphasis detection accuracy on similar data [27].

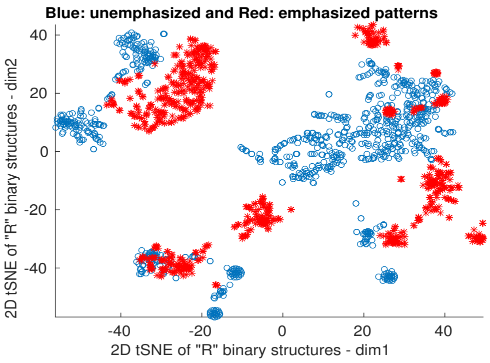

Figure 3: *tSNE visualization of different realizations of phoneme "R" with and without emphasis.*

4. Conclusions and Future Directions

Posterior exemplars are sparse and live in union of low-dimensional subspaces. In this paper, a novel application of this property was proposed by introducing the quantized posterior hashing technique to enable an effective search strategy confined to the local neighborhood of posterior exemplars. The quantized posteriors are regarded as the representatives of the posterior space. It was shown that the number of unique hash keys or equivalently the number of buckets is very small that leads to tremendous reduction of the search space with negligible overhead. This method enables very fast and accurate kNN search for phonetic classification and linguistic parsing.

The number of unique hash keys or different buckets is related to the sparsity level or the number of permissible classes. Since the "optimal" phonetic posteriors indicate a single highly probable class, the number of buckets is proportional to the number of phones. In contrast, the phonological posteriors are indicative of sub-phonetic attributes presented at multiple phonological class probabilities, hence, the number of unique hash keys is proportional to the size of permissible combinations as roughly quantified at binary quantization. In fact, binary level hashing is very efficient for neighborhood search of phonological posterior exemplars.

Future work will focus on development of fast and flexible large scale ASR in a hierarchical exemplar based framework relying on the generic low-dimensional properties of DNN posterior exemplars. Furthermore, we will investigate alternative linguistic parsing tasks using structured sparsity of phonological posteriors. Higher semantic information can be integrated with the bottom-up approach towards development of exemplar based ASR framework that can exploit broad contextual information. Moreover, the neural network posterior estimation can be seen as a hash function of any type of speech acoustic features such as spectral features. Hence, exemplar-based speech classification can potentially benefit from the proposed quantized posterior hashing at any speech representation level. We will study this idea in future developments.

5. Acknowledgments

The research leading to these results has received funding from by SNSF project on "Parsimonious Hierarchical Automatic Speech Recognition (PHASER)" grant agreement number 200021-153507.In addition, this work has been conducted with the support of the Swiss NSF under grant SP2: the SCOPES Project on Speech Prosody.

6. References

[1] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice-Hall London, 1982, vol. 761.

[2] Y. Xu, O. Siohan, D. Simcha, S. Kumar, and H. Liao, "Exemplar-based large vocabulary speech recognition using k-nearest neighbors," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2015, pp. 5167–5171.

[3] N. Singh-Miller, *Neighborhood Analysis Methods in Acoustic Modeling for Automatic Speech Recognition*. MIT PhD Thesis, 2010.

[4] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2067–2080, 2011.

[5] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4370–4373.

[6] J. Wang, H. T. Shen, J. Song, and J. Ji, "Hashing for similarity search: A survey," *CoRR*, vol. abs/1408.2927, 2014. [Online]. Available: http://arxiv.org/abs/1408.2927

[7] M. Slaney and M. Casey, "Locality-sensitive hashing for finding nearest neighbors [lecture notes]," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 128–131, 2008.

[8] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, "Exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[9] G. Luyet, *Low-Rank Representation for Enhnaced Enhanced Deep Neural Network Acoustic Modeles*. Idiap Research Intitute Master Thesis, 2016.

[10] P. Dighe, A. Asaei, and H. Bourlard, "Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition," *Speech Communication*, 2015.

[11] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 421–426.

[12] D. Ram, A. Asaei, P. Dighe, and H. Bourlard, "Sparse modeling of posterior exemplars for keyword detection," in *Proceedings of Interspeech*, no. EPFL-CONF-209088, 2015.

[13] D. Ram, A. Asaei, and H. Bourlard, "Sparse subspace modeling for query by example spoken term detection," Idiap, Idiap-RR Idiap-RR-01-2016, 1 2016.

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

[15] ——, "On structured sparsity of phonological posteriors for linguistic parsing," *arXiv preprint arXiv:1601.05647*, 2016.

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

[17] 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

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

[19] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction (MLMI)*, 2006, pp. 28–39.

[20] 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*, Dec. 2011.

[21] I. Himawan, P. Motlicek, M. Ferras, and S. Madikeri, "Towards utterance-based neural network adaptation in acoustic modeling," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, no. EPFL-CONF-213067, 2015.

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

[23] 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, 2006.

[24] "Spoken interaction with interpretation in switzerland (SIWIS)," 2015. [Online]. Available: https://www.idiap.ch/project/siwis/downloads/siwis-database

[25] A. Asaei, H. Bourlard, and B. Picart, "Investigation of knn classifier on posterior features towards application in automatic speech recognition," Idiap-RR Idiap-RR-11-2010, 6 2010.

[26] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[27] M. Cernak and P. Honnet, "An empirical model of emphatic word detection," in *16th Annual Conference of the International Speech Communication Association INTERSPEECH*, 2015, pp. 573–577.