Interspeech 2018
2-6 September 2018, Hyderabad

Phonological Posterior Hashing for Query by Example Spoken Term Detection

Afsaneh Asaei[†], Dhananjay Ram[†,‡], Hervé Bourlard[†,‡]

[†]Idiap Research Institute, Martigny, Switzerland
[‡]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{dhananjay.ram, afsaneh.asaei, herve.bourlard}@idiap.ch

Abstract

State of the art query by example spoken term detection (QbE-STD) systems in zero-resource conditions rely on representation of speech in terms of sequences of class-conditional posterior probabilities estimated by deep neural network (DNN). The posteriors are often used for pattern matching or dynamic time warping (DTW). Exploiting posterior probabilities as speech representation propounds diverse advantages in a classification system. One key property of the posterior representations is that they admit a highly effective hashing strategy that enables indexing a large audio archive in divisions for reducing the search complexity. Moreover, posterior indexing leads to a compressed representation and enables pronunciation dewarping and partial detection with no need for DTW. We exploit these characteristics of the posterior space in the context of redundant hash addressing for query-by-example spoken term detection (QbE-STD). We evaluate the QbE-STD system on AMI corpus and demonstrate that tremendous speedup and superior accuracy is achieved compared to the state-of-the-art pattern matching solution based on DTW. The system has the potential to enable massively large scale spoken query detection.

Index Terms: Posterior probability structures, Posterior hashing, Pronunciation dewarping, Structural similarity measure, Query by example, Spoken term detection.

1. Introduction

Query-by-Example Spoken Term Detection (QbE-STD) refers to the task of finding out audio documents containing a spoken query. The query is uttered by the user so only very few (or just a single) examples are provided. This system enables the user to search over multi-lingual audio archives without any prior assumption on the language of the query. Hence, the task is inherently language independent and no (or quite limited) linguistic resources may be available for system development.

1.1. State-of-the-art Solutions and Challenges

QbE-STD received serious consideration in the context of MediaEval spoken query search benchmarking campaign [1, 2, 3]. Recent exemplar based speech processing offers high flexibility in speech applications, partly attributed to the lack of complex statistical assumptions that facilitate exploiting "data deluge" with no prejudice on expected answers. Deep neural network (DNN) based class-conditional posterior probabilities (hereafter referred to as *posteriors*) have been found to be one of the best speech representations to enable exemplar based speech recognition [4] and spoken query detection [5, 6, 7]. In theory, if infinite number of exemplars of continuous probability density functions are provided, a simple nearest-neighbor rule leads to optimal classification [8].

Nevertheless, exemplar-based speech processing faces two fundamental problems: (1) The growing size of the databases prohibits efficient search, and (2) The duration variation in speech pronunciation is effectively handled via dynamic time warping that is computationally expensive and sub-optimal due to dependency on the local reference exemplar. This paper addresses these limitations to foster exemplar based solutions for real time applications.

DNN posteriors live in union of low-dimensional structured sparse subspaces [9,10]. Exploiting this property enables a hierarchical speech classification and recognition framework based on structured sparse modeling of posterior exemplars [9, 11]. In addition, the low-dimensional subspaces can be modeled through dictionary learning for sparse coding to enable unsupervised adaptation and enhanced acoustic modeling for speech recognition [10, 12]. Sparse subspace modeling of the posterior exemplars are also found promising for query-by-example spoken term detection (QbE-STD) [7, 11, 13].

Recently, we investigated a novel application of structured sparsity of posterior probabilities in devising an effective hashing technique to reduce the search space of posterior exemplars [14]. Application of hashing in exemplar search enables splitting the search space into disjoint buckets each indexed with a unique hash key (posterior representative). The exhaustive search space is thus downsized to the corresponding bucket sizes. In this context, the hash function ensures geometric locality preserving of neighboring examples [14, 15]. In this paper, we propose a highly efficient QbE-STD system exploiting posterior hashing. The framework is inspired from the idea of redundant hash addressing.

Redundant hash addressing (RHA) was initially proposed by Teuvo Kohonen as a fast method for recognition and correction of garbled symbol strings. It is an associative method based on the use of multiple (redundant) features extracted from the same input item [16]. The comparison of the input item against the reference items is based on these features. Redundancy is exploited to increase error tolerance and robustness. Kohonen applied this idea for word recognition. The segments of N consecutive letters (N-grams) are considered. The RHA system consists of the N-gram table and word dictionaries. Multiple features (N-grams) are extracted from the input string and each extracted N-gram associates the input string with a word in the dictionary based on the number of matching N-grams [16].

To obtain the N-grams of symbols/letters for RHA, the acoustic feature vectors are first quantized and mapped into a symbol space. To that end, the self organizing map (SOM) neural network is used as a codebook to map the input feature vectors into the finite set of prototype vectors. When each prototype vector is provided with an index, feature vector sequence can be mapped into a symbolic index sequences. Each feature vector is encoded by the index of its best matching unit. The node indices of the SOM are thus the alphabet of the system [17]. In this paper, we propose that posterior hashing can be used to define the codebook for RHA.

10.21437/Interspeech.2018-1973

In the following Section 2, we briefly explain the idea of posterior hashing. The new framework of RHA for QbE-STD is explained in Sections 3. We also exploit posterior hashing to develop an efficient pattern matching method in Section 4. The experiments are conducted in Section 5, and the conclusions are drawn in Section 6.

2. Hashing Structured Posterior Probabilities

We consider the posterior vector consisting of Q class-conditional posterior probabilities estimated by DNN from the input acoustic feature $\boldsymbol{x}$, denoted as

$$\boldsymbol{z} = [p(C_1|\boldsymbol{x}), \ldots, p(C_q|\boldsymbol{x}), \ldots, p(C_Q|\boldsymbol{x})]^\top \quad (1)$$

where $.^\top$ is the transpose operator. The posteriors can be defined at any linguistic level. The typical phone and phonological posteriors are shown to be highly structured and living in low-dimensional subspaces [12, 18]. Taking advantage of the underlying structured sparsity of posteriors, a hashing technique is devised to divide the space into smaller size buckets of neighboring posteriors based on the following hashing formula

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

where b is the number of bits for quantization. The number of unique quantized posteriors is 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 as disjoint buckets.

In theory, quantization of every component of posteriors in b bits leads to splitting the space in maximum 2^K disjoint regions where $K = 2^b Q$. Accordingly, the size of training data in each bucket can be reduced to an average $N/2^K$. The analysis in [14] shows that the probability of negative examples in a bucket is $1 - 2^{-Kb}$. Considering the typical value of Q for phonetic or phonological posteriors, this hashing function leads to a very small probability of encapsulating negative examples or wrong positive examples in the same bucket. In practice, the quantized posterior hashing is found to reduce the search space drastically with no degradation in performance.

Inspired from the idea of redundant hash addressing (RHA) for recognition of word sequences, we revisit its implications and applicability for QbE-STD task. This application provides a unique case study where the potential of posterior hashing is fully exploited to speed up the query search on large speech archives. In the following section, we review the principles of RHA for query detection.

3. Redundant Hash Addressing

We adopt the basic word recognition RHA framework for query detection. The dictionary consists of the query term. The speech utterances are represented in terms of posteriors estimated for short frames. The posteriors are converted into codes exploiting quantized posterior hashing. To that end, the training data is quantized and the unique binary codes form the codebook of the hash keys or symbols in a hash space.

The dynamics of speech production is slower than the short frame sampling frequency. Therefore, adjacent frames are likely to share similar codes. To obtain a *duration-invariant* representation, the similar codes of adjacent frames are merged. This approach enables an efficient method to deal with the duration variations in spoken utterances and queries.

Once the testing utterances and the spoken queries are converted into this code space, N-grams are formed by concatenating each code with $N-1$ adjacent ones on its right. Then, the N-grams of the utterance and the query are compared. If a matching code is occurred, it is labeled as 1 and 0 otherwise. In this procedure, the N-grams capture the trajectory information and they are processed independently. The number of detected N-grams is used as the score for query detection. Fig. 1 illustrates the RHA framework for query detection.

Figure 1: *Building blocks of the RHA based QbE-STD system: The sequence of posteriors is mapped to a sequence of symbols each associated to a unique hash code. The N-grams of the query dictionary are matched against the N-grams of the spoken utterance. The number of matching N-grams is used as the score for query detection.*

Alternative solutions for detection tasks using posteriors rely on nearest neighbor approaches to pattern matching [19]. Hence, we investigate this idea and exploit posterior hashing to develop an efficient pattern matching framework in the following section.

4. Structural Pattern Matching

One key problem of speech pattern matching is handling the duration variation in speech production. Posterior hashing can be used for segmentation of the similar codes in a pronunciation dewarping mechanism.

4.1. Dewarping for Duration Invariance

The benefit of the redundant hash addressing principle is that the duration variation can be addressed at the *representation* level rather than the *recognition* level through DTW. We exploit the *duration invariant* representation enabled by hashing for segmentation of the speech utterance. To that end, *blocks* of similar hash keys are identified. It is hypothesized that the posteriors encapsulated in a block represent temporal duration of a discrete production process. This idea was previously found effective in duration analysis of impaired speech production [20]. In this work, we use the blocking procedure to address the duration variation in speech representation. The posteriors of a block are averaged to form an average pronunciation. The crucial factor in the pattern matching system is the choice of similarity measure.

4.2. Structural Similarity Measure

The posterior space is highly structured and low-dimensional [18, 21]. To exploit this property, we propose to use Spearman's rank correlation to measure the similarity of posterior exemplars. The intuition is that the exact value of the posteriors is less important compared to the structure of the high probability components. The high probability components quantify the order of significance in structuring the speech

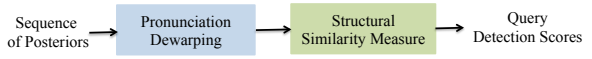

Figure 2: *Building blocks of the structural pattern matching for QbE-STD: The sequence of posteriors is processed in blocks according to the similar hash keys. The blocked posteriors are replaced with the average posterior to obtain a duration invariant representation. Structural similarity of the posteriors is measured via Spearman's similarity measure* (3)*. The similarity scores are integrated based on max-sum dynamic programming to obtain the query detection score.*

signal. The Spearman's similarity measure is defined as

$$S_{\text{Spearman}}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \frac{(\boldsymbol{r}_{\boldsymbol{z}_1} - c)(\boldsymbol{r}_{\boldsymbol{z}_2} - c)^\top}{\sqrt{(\boldsymbol{r}_{\boldsymbol{z}_1} - c)(\boldsymbol{r}_{\boldsymbol{z}_1} - c)^\top}\sqrt{(\boldsymbol{r}_{\boldsymbol{z}_2} - c)(\boldsymbol{r}_{\boldsymbol{z}_2} - c)^\top}}. \quad (3)$$

where $\boldsymbol{r}_{\boldsymbol{z}_1}$ and $\boldsymbol{r}_{\boldsymbol{z}_2}$ are the coordinate-wise rank vectors of the posterior vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, and $c = \frac{Q+1}{2}$. The Spearman similarity and the coordinate-wise rank vectors are computed using MATLAB.

The frame level similarity scores constitute a curve indicating the probability of a query occurring in a test utterance. We use max-sum dynamic programming to obtain a region of occurrence and the corresponding area under the curve is used as the score for query detection [7]. This procedure is illustrated in Fig. 2.

5. Experimental Evaluation

The experiments are conducted to evaluate the performance of the proposed methods in challenging scenarios when just one query example is provided for QbE-STD, and the query and background are conversational spontaneous speech with interfering speakers.

5.1. Benchmarking Setup

The AMI meeting corpus (IHM) [22] is used for the experiments where the training, development and evaluation sets are as [23]. Although the meeting language was English, many participants were non-native speakers. Also, the headset recordings contain considerable amount of overlapping speech due to interfering speakers. There are approximately 12 k words in the training, out of which 100 words are randomly used for our detection experiments including very short words such as "ten" to long words such as "requirements". The 9 hours of speech in the evaluation set is used as the search space for QbE-STD. The total number of search utterances for query detection is 10179. The average number of positive examples for all queries is 46 where the number of positive examples per query varies between 3 to 273 with a standard deviation of 52. A single query example is chosen randomly from the training set for query detection and it is used for all the systems.

5.2. Baseline System

The DTW based QbE-STD system presented in [1] is used as a competitive baseline system [2]. It consists of following steps. First, posterior features are extracted from both spoken query and test utterance. These features vectors are then used to compute a frame-level distance matrix using cosine similarity. A modified DTW algorithm is employed to find a warping path through this matrix and compute the corresponding likelihood score. This DTW approach is similar to slope constrained DTW [5] where the optimal warping path is normalized by its partial path length at each step and constraints are imposed so that the warping path can start and end at any point in the test utterance.

5.3. Posterior Representation

In order to obtain posterior representation of the data, we consider phonological posteriors. We use the open-source DNN based phonological vocoding platform [24] for estimation of the extended Sound Pattern of English (eSPE) phonological posteriors. The motivation for using phonological posteriors is three-fold: (1) Phonological posterior quantization and hashing is found to be effective in search space reduction for accurate classification [14, 18, 21], (2) Sub-phonetic nature of phonological posteriors facilitates development of flexible and low-resource speech detection and recognition solutions [25], and (3) Phonological posterior are found robust for inter-domain posterior representation where the training and testing acoustic conditions and languages are different [14, 18, 21].

The ultimate goal of a large-scale QbE-STD system is to operate on non-native speech of multiple languages across diverse domains of speech recordings. Hence, the training data of AMI is not used and the DNN setup is trained on the Wall Street Journal (WSJ) continuous speech recognition corpora [26]. It consists of 21 different DNNs corresponding to each phonological class including one class for silence. All DNNs have 3 hidden layers of 1024 neurons per layer. They were trained using mel frequency cepstral coefficient feature (MFCC) with a context of 9 frames. The 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 vs the rest. The outputs of all DNNs are concatenated to form a phonological posterior vector [24].

5.4. QbE-STD Results

The QbE-STD evaluation results are illustrated in Fig. 3 where $N = 2$ is considered for N-grams used in RHA. We can see that RHA is the best performing system. Previous studies show that posterior classification is most accurate when cosine similarity is used [1,27]. However, we can see that pattern matching using continuous posterior features is more effective when structural similarity is exploited. The Spearman similarity yields up to 5% reduction in the miss-rate at most operating points.

We observed no degradation in pattern matching performance due to dewarping. This result was expected due to the binary nature of phonological posteriors [18, 21]. Moreover, if the dewarped posteriors are quantized into binary vectors and Jaccard similarity is used for binary pattern matching [21], similar results as the Spearman similarity measure is achieved. This observation again confirms that the space of phonological posteriors is highly structured and the structures bear more information than the exact posterior values.

Each component of a phonological posterior indicates the probability of a phone attribute composing a phonetic unit [24]. The permissible combinations are highly constrained due to articulatory mechanisms governing speech production. Therefore, the probabilities constituting a posterior are confined to a small number of components where the indices of high probabilities determine the unique structure of the vocal machinery in speech production [21].

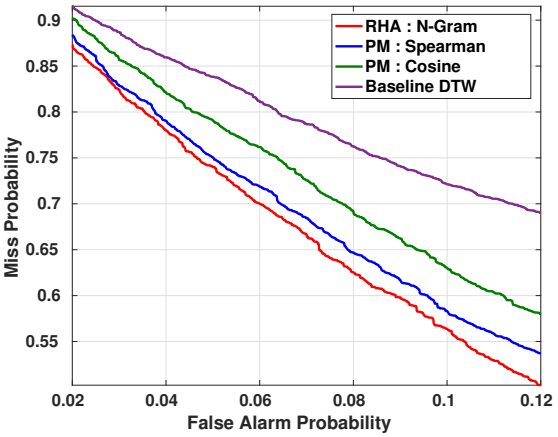

Figure 3: *QbE-STD performance on AMI database using (i) redundant hash addressing (RHA, Section 3) with N-Gram matching, (ii) Pattern matching (PM, Section 4) with Spearman structural similarity measure, and (iii) Pattern matching with cosine distance to measure posterior similarity [1, 27]. The DET curves are compared with a competitive baseline DTW [1, 2] system. A single example is used per query for detection.*

5.5. Search Reduction

Posterior hashing reduces the computational cost through (1) reducing the search space to a compressed space of unique codes and (2) reducing the cost of similarity measure computation to a look-up table associated to matching codes.

In the case of redundant hash addressing, the size of test data is reduced to 0.0037 of the initial number of frames. The size of query exemplars is also reduced to 0.4 of the initial size. Hence, the search space is reduced to 0.0015 or nearly 10^4-fold reduction. More precisely, the size of the AMI test set is 2500333 frames that is reduced to 9216 unique codes exploiting posterior hashing. It may be noted that the number of unique codes obtained from phonological posteriors extracted for AMI corpus is only 0.0044 of the total number of possible codes (2^Q), where Q denotes the number of phonological classes that is 21 in this study.

In the case of pattern matching, application of hashing reduces the search space of exemplars to 0.16 of the initial size. More concretely, the number of test frames is reduced by 0.4 and similarly, the number of query frames is also reduced to an average 0.4 of the original size. Both RHA and pattern matching can exploit binary pattern matching and it can be implemented efficiently through a look-up table of code distances in an offline preparation.

In general, assuming N number of frames, the computational complexity of DTW is $O(N^2)$. The proposed pattern matching has the complexity of $O(N)$ where N is effectively reduced using the dewarping procedure. The computational complexity of the RHA is $O(C)$ where C denotes the number of unique binary codes. It may be noted that C depends on the pronunciation variations and it does not grow with N due to the increasing number of similar pronunciations in growing size of the speech archives.

6. Conclusions

Speech representation in terms of posterior probabilities offers diverse benefits for speech classification applications, in particular solutions relying on exemplar matching. Posterior representations are highly structured and low-dimensional. We exploit this property in devising an effective hashing technique to define data driven symbols or codes. Redundant hash addressing is applied on the posterior codes to enable fast query search by detecting the matching codes. A fast QbE-STD system is achieved where the search space is reduced by a factor of 10^4. The system is compared to a state-of-the art DTW based pattern matching algorithm and it outperforms the alternative slower solution. The unique codes encapsulate the structure of pronunciations and their number is expected to be confined to a small number of permissible articulatory structures regardless of the growing size of the speech databases. Hence, redundant hash addressing incorporating posterior hashing can lead to highly efficient solutions for massively large scale query search. We plan to investigate language independent QbE-STD system development in future studies.

7. Acknowledgments

We acknowledge Dr. Milos Cernak for helping with the phonological posteriors estimation. The research leading to these results has received funding from by Swiss NSF project on "Parsimonious Hierarchical Automatic Speech Recognition and Query Detection (PHASER-QUAD)" grant number 200020-169398.

8. References

[1] L. J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, "High-performance query-by-example spoken term detection on the SWS 2013 evaluation," in *Proceedings of ICASSP*, 2014, pp. 7819–7823.

[2] X. Anguera, L. J. Rodriguez-Fuentes, I. Szöke, A. Buzo, and F. Metze, "Query by example search on speech at mediaeval 2014." in *MediaEval*, 2014.

[3] I. Szöke, L. J. Rodríguez-Fuentes, A. Buzo, X. Anguera, F. Metze, J. Proenca, M. Lojka, and X. Xiong, "Query by example search on speech at mediaeval 2015." in *MediaEval*, 2015.

[4] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1377–1390, 2007.

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

[6] 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.

[7] D. Ram, A. Asaei, and H. Bourlard, "Sparse subspace modeling for query by example spoken term detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1130–1143, June 2018.

[8] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*, 1982, vol. 761.

[9] D. Ram, A. Asaei, P. Dighe, and H. Bourlard, "Sparse modeling of posterior exemplars for keyword detection," in *Proceedings of Interspeech*, 2015.

[10] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, "Exploiting low-dimensional structures to enhance DNN based acoustic modeling in speech recognition," in *Proceedins of ICASSP*, 2016.

[11] D. Ram, A. Asaei, and H. Bourlard, "Subspace detection of dnn posterior probabilities via sparse representation for query by example spoken term detection," in *Proceedings of Interspeech*, 2016.

[12] G. Luyet, P. Dighe, A. Asaei, and H. Bourlard, "Low-rank representation of nearest neighbor phone posterior probabilities to enhance DNN acoustic modeling," in *Proceedings of Interspeech*, 2016.

[13] D. Ram, A. Asaei, and H. Bourlard, "Subspace regularized dynamic time warping for spoken query detection," in *Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2017.

[14] A. Asaei, G. Luyet, M. Cernak, and H. Bourlard, "Phonetic and phonological posterior search space hashing exploiting class-specific sparsity structures," in *Proceedings of Interspeech*, 2016, pp. 1873–1877.

[15] 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.

[16] T. Kohonen, H. Riittinen, M. Jalanko, E. Reuhkala, and S. Haltsonen, "A thousand-word recognition system based on the learning subspace method and redundant hash addressing," in *Proceedings of the 5th International Conference on Pattern Recognition*, vol. 1, 1980, pp. 158–165.

[17] P. Somervuo, "Redundant hash addressing of feature sequences using the self-organizing map," *Neural processing letters*, vol. 10, no. 1, pp. 25–34, 1999.

[18] 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.

[19] Y. Zhang, "Unsupervised speech processing with applications to query-by-example spoken term detection," Ph.D. dissertation, Massachusetts Institute of Technology, 2013.

[20] A. Asaei, M. Cernak, and M. Laganaro, "PAoS markers: Trajectory analysis of selective phonological posteriors for assessment of progressive apraxia of speech," in *Proceeding on the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016.

[21] M. Cernak, A. Asaei, and H. Bourlard, "On structured sparsity of phonological posteriors for linguistic parsing," *Speech Communication*, vol. 84, pp. 36–45, 2016.

[22] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[23] "AMI corpus partition," http://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml.

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

[25] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

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

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