COMBINING EXEMPLAR-BASED MATCHING AND EXEMPLAR-BASED SPARSE REPRESENTATIONS OF SPEECH

Emre Yılmaz, Dirk Van Compernolle, and Hugo Van hamme

Dept. ESAT, KU Leuven, Leuven, Belgium

ABSTRACT

In this paper, we compare two different frameworks for exemplar-based speech recognition and propose a combined system that approximates the input speech as a linear combination of exemplars of variable length. This approach allows us not only to use multiple length long exemplars, each representing a certain speech unit, but also to jointly approximate input speech segments using several exemplars. While such an approach is able to model noisy speech, it also enforces a feature representation in which additivity of the effect of signal sources holds. This is observed to limit the recognition accuracy compared to e.g. discriminatively trained representations. We investigate the system performance starting from a baseline single-neighbor exemplar matching system using discriminative features to the proposed combined system to identify the main reasons of recognition errors. Even though the proposed approach has a lower recognition accuracy than the baseline, it significantly outperforms the intermediate systems using comparable features.

Index Terms— speech recognition, exemplar-based, template matching, sparse representations

1. INTRODUCTION

Exemplar-based (or template-based) speech recognition recently regained popularity due to the significant increase in computational power and development of fast template matching and search algorithms [1]. Several hybrid recognition systems combining this approach with hidden Markov models (HMMs) are also proposed [2, 3]. Exemplars are labeled speech segments such as phones or syllables, possibly of different length, that have occurred in the training data and they are matched with the input speech segments using dynamic time warping (DTW). We refer to this approach as *exemplar matching*. This approach allows to use any choice of frame-synchronous feature vector to represent the input speech and the exemplars. For instance, in [1], motivated by a better recognition accuracy, a mutual information based discriminant analysis (MIDA [4]) is applied to log-spectral data.

One can simply classify the segment as the label of the closest exemplar, or by a voting scheme on the set of K nearest neighbors [1, 5]. Applying exemplar matching under noisy conditions creates mismatch problems similar to what is experienced with HMMs. One can resort to feature compensation methods to increase the robustness to noise [6]. Model compensation techniques require would require all exemplars to be modified, which is a formidable task in the case of non-stationary noise. Since the search problem in exemplar-based recognition is a lot more involved than in HMM-based recognition, the equivalent of factorial models is also not a trivial path to walk. Finally, multi-condition training, i.e. storing noisy exemplars, will increase the number of exemplars dramatically. Furthermore, noisy exemplars can only capture a certain instant of speech and noise resulting in a limited noise modeling especially in case of non-stationary noise.

More recently, exemplar-based *sparse representations* have been used successfully for clean [7, 8] and noisy [9, 10, 11] speech recognition. This technique models input speech segments as a sparse linear combination of fixed-length exemplars. These exemplars are represented in the linear magnitude spectral domain to ensure additivity. By combining speech and noise exemplars linearly, it explicitly models the noisy speech. Because exemplars are combined linearly, they need to be of the same length, unlike in exemplar matching, and cannot model our choice of speech segments (phones, syllables, ...). The exemplars can therefore not serve directly as an acoustic model, so sparse representations have been used for speech enhancement, a model of state likelihoods (sparse classification) or to generate a mask in a missing data recognition framework.

In this paper, we elaborate on the differences between the DTW and sparse representation exemplar techniques and propose a procedure to combine them. This results in a basic exemplar matching recognizer having the advantage of using long exemplars of variable length in a sparse representation formulation. The main motivation is to establish a new framework that allows noise modeling for exemplar matching based recognition systems. This task involves both the selection of the appropriate representation domain of speech and the distance/divergence measure used for comparing the input speech segments with exemplars. Most exemplar matching techniques make use of state-of-the-art features with high discriminative power among the classes to lower the recognition errors [1, 5]. However, as additivity and non-negativity properties are required for linearly combining exemplars, mel-scaled magnitude and power spectra can be used to represent speech in the proposed approach. The Euclidean distance used in exemplar matching has to be replaced by e.g. the generalized Kullback-Leibler divergence. This study focusses on the price that needs to be paid in terms of the accuracy on *clean* data for these modifications. An analysis of the resulting noise robustness is the topic of other work currently under review.

The rest of the paper is organized as follows. Section 2 explains exemplar matching based recognition, exemplar-based sparse representations of speech and the combined system. The experimental setup is discussed in Section 3. Section 4 presents the results. The conclusions are discussed in Section 5.

2. EXEMPLAR-BASED RECOGNITION SYSTEMS

2.1. Exemplar-matching

This technique compares the input speech segments with labeled exemplars, each representing a certain class. The exemplars are collected from a large corpus that is segmented in terms of the desired classes. The segments will have variable lengths, so the natural du-

ration distribution of each class in the training corpus is preserved. Input speech and exemplars are represented using state-of-the-art speech features in order to maximize recognition accuracy. Recognition then consists of finding the sequence of exemplars that best matches the input subject to lexical and grammatical segment concatenation constraints. The quality of a match is measured by a metric (e.g. Euclidean distance) that expresses how well the exemplars reconstruct the data. Additional constraints are imposed. Each exemplar is tagged with meta-information such as speaker characteristics (e.g. gender, age) or prosodic information (e.g. speaking rate, position in the sentence). This information is used during decoding to penalize inconsistent exemplar sequences (e.g. mixed gender) with various concatenation costs. In the present work, only two types of concatenation costs are considered, namely exemplar startup costs and gender costs. Exemplar startup costs penalize longer exemplar sequences and control the insertion/deletion rate. Gender costs penalize mixed gender exemplar sequences, a constraint which has been shown to improve the recognition accuracy [1]. Finally, in earlier exemplar matching work, strict matches across the time dimension were relaxed using DTW. In this work, time warping is not applied for three reasons. Firstly, it would complicate the distance calculation. Secondly, in noisy conditions, too much freedom in time warping may lead to unrealistic warping, so duration constraints are more important than in clean conditions. The same effect has been observed in HMM systems [12]. Thirdly, in the combined system described in Section 2.3, the linear combination of exemplars with different internal time warping will relax the requirement for strict matching along the time axis.

2.2. Sparse Combinations of Exemplars

The exemplar-based sparse representations approach models the input speech as a linear combination of several speech exemplars [7]. The input speech and exemplars are represented in the linear mel-scaled spectral domain in order to ensure additivity of exemplars. In this framework, exemplars are fixed-length speech segments randomly extracted from the training corpus and may be associated with more than one class. Labeling is performed probabilistically using a conventional HMM-based recognizer either at the word or state level.

Exemplars consisting of L frames are reshaped as a single column vector and collected in a single dictionary $\mathbf{S}$ of dimensionality $DL \times N$ where D is the number of frequency bands and N is the number of available exemplars. A reshaped input speech vector $\mathbf{y}_L$ of length L is expressed as a linear combination of the exemplars with non-negative weights:

$$\mathbf{y}_L \approx \sum_{m=1}^{N} x_m \mathbf{s}_m = \mathbf{S}\mathbf{x} \qquad \text{s.t.} \qquad x_m \geq 0 \tag{1}$$

where $\mathbf{x}$ is an N-dimensional sparse weight vector. Sparsity of the weight matrix implies that the input speech is approximated by a small number of exemplars. The exemplar weights are obtained by minimizing the cost function,

$$d(\mathbf{y}_L, \mathbf{S}\mathbf{x}) + \sum_{m=1}^{N} x_m \Lambda_m \qquad \text{s.t.} \qquad x_m \geq 0 \tag{2}$$

where $\mathbf{\Lambda}$ is an N-dimensional vector. The first term is the divergence between the input speech vector and its approximation. A regularization term is added in order to limit the l_1-norm of the weight vector.

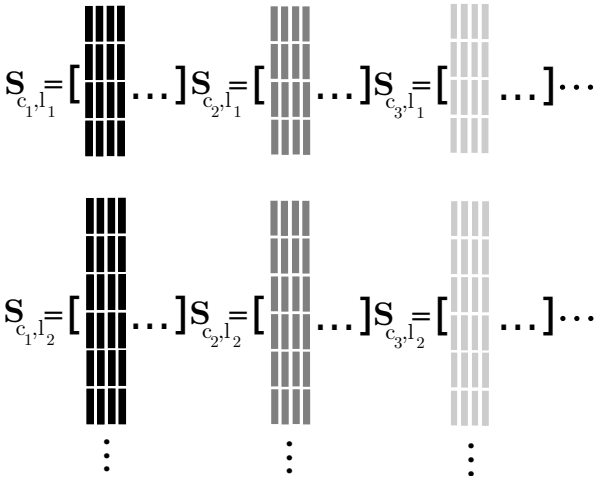

Fig. 1. Exemplars are organized in multiple dictionaries $\mathbf{S}_{c,l}$ for each class c and each length l.

Here, $\mathbf{\Lambda}$ controls how sparse the resulting vector $\mathbf{x}$ is. The generalized Kullback-Leibler divergence (KLD) is used for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k \tag{3}$$

which is commonly used in source separation problems and shown to produce better results than Euclidean distance when used with linear mel-scaled spectra [13].

The regularized convex optimization problem can be solved using various methods including LASSO and non-negative sparse coding (NSC). For NSC, the multiplicative update rule to minimize the cost function (2) is derived in [14] and found as

$$\mathbf{x} \leftarrow \mathbf{x} \odot (\mathbf{S}^T(\mathbf{y}_L \oslash (\mathbf{S}\mathbf{x}))) \oslash (\mathbf{S}^T\mathbf{1} + \mathbf{\Lambda}) \tag{4}$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}$ is a DL-dimensional vector with all elements equal to one. Applying this update rule iteratively, the weight vector becomes sparse and the reconstruction error between the input speech vector and its approximation decreases monotonically.

In order to decode the input speech, a window of length L is slid over the input speech with a constant frame shift and the weight vector for each window is obtained. Then, using a label matrix containing the word or state based labels for each exemplar, the HMM likelihood scores are calculated. Finally, a modified Viterbi algorithm is applied to find the most likely class sequence.

2.3. Combined System

The combined system aims to benefit from the advantages of the two frameworks explained in the previous sections. It is an exemplar matching approach in the sense that it explains the input as the sequence of classes leading to a minimal reconstruction error, each class being represented by exemplars of variable length. The reconstruction error is however measured by the sparse combination model in the linear spectral domain, which has the advantage of easily modeling noisy speech by adding noise exemplars. The exemplars are thus organized in multiple dictionaries $\mathbf{S}_{c,l}$: one for each class c and each length l as shown in Figure 1. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where $N_{c,l}$ is the number of available exemplars of length l and class c. Using separate dictionaries for different classes is expected to provide better classification than using a single dictionary as every input segment is guaranteed to be

approximated by a combination of exemplars belonging to the same class only.

For any class c, a reshaped input speech vector $\mathbf{y}_l$ of length l is expressed as a linear combination of the exemplars with non-negative weights:

$$\mathbf{y}_l \approx \sum_{m=1}^{N_{c,l}} x_{c,l}^m \mathbf{s}_{c,l}^m = \mathbf{S}_{c,l}\mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \qquad (5)$$

where $\mathbf{x_{c,l}}$ is an $N_{c,l}$-dimensional sparse weight vector. The class and length dependent weight vectors are obtained by applying the multiplicative update rule in Equation (4) for each dictionary. The reconstruction error between a class c and an input speech segment of length l can be calculated using Equation (3). It satisfies the conditions to apply dynamic programming, hence the class sequence that best matches the input speech can be simply found.

The input speech is decoded similar to the exemplar matching based recognizer. Every input frame sequence of each available exemplar length is approximated as a linear combination of exemplars by iteratively applying the update formula. For each class and exemplar length, the approximation is performed separately using the dictionaries. After a certain number of iterations, the reconstruction error is calculated using Equation (3). As every dictionary contains exemplars with known labels, the entire input utterance is searched to find the digit sequence yielding the minimum reconstruction error.

A known problem of sparse representation approaches working on magnitude spectra is that the silence exemplars are not recognized [14]. This is due to the fact that silence is well-approximated by combining speech exemplars with small weights, so all classes will score equally well. To overcome this problem, reconstruction errors for the class representing silence have to be compensated. This is achieved by reducing the reconstruction errors corresponding to silence dictionaries by a compensation factor CF depending on the voice activity value assigned to the middle frame of the corresponding input speech segment and the reconstruction error itself,

$$CF = C \cdot d(\mathbf{y}_l, \mathbf{S}_{sil,l}\mathbf{x}_{sil,l}) \cdot VAD \qquad (6)$$

where C is a scale factor and VAD is the voice activity estimate (0 for speech, 1 for silence). The VAD value can either be obtained from an autonomous module implementing a preferred method from the vast literature on the topic, or it can be estimated using the exemplar weights $\mathbf{x_{c,1}}$. In this work, an energy-based VAD is used. It should be noted that including the reconstruction error itself in Equation (6) compensates for length differences.

3. EXPERIMENTAL SETUP

3.1. Data, Preprocessing and Features

We have conducted recognition experiments on the 4 clean test sets of the AURORA-2 database [15]. To reduce simulation time, we subsampled each test set by a factor of 4, bringing the total number of utterances to 1001. For feature extraction, a 17 channel Mel-scaled filter bank with triangular magnitude response is computed from a spectral analysis with a window length of 32 ms and a frame shift of 10ms. The first channel is centered at 200 Hz and the last at 3030 Hz. Channel normalization of the magnitude spectrum is achieved by transforming it to the log-domain, applying mean normalization and moving back to the linear domain. The exemplar matching baseline uses MIDA features, i.e. a discriminatively trained linear transform of the mean-normalized log-power spectra and its first and second order differences (a total of $3 \times 17 = 51$ features) resulting in 32-dimensional feature vectors.

Table 1. Word error rates for the 1-NN exemplar matching based recognizer in percentages

Features	Dimension	Dis./Div. Measure	WER (%)
MIDA	32	Eucl.	1.11
MN+logPowSpec	17	Eucl.	3.36
MN+PowSpec	17	KLD	10.10
MN+MagSpec	17	KLD	4.41
PowSpec	17	KLD	10.34
MagSpec	17	KLD	4.36

Table 2. Word error rates for the proposed system in percentages

Features	Dimension	Dis./Div. Measure	WER (%)
PowSpec	17	KLD	7.70
MagSpec	17	KLD	2.98
l_2-N+PowSpec	17	KLD	5.14
l_2-N+MagSpec	17	KLD	2.16

3.2. Exemplar Matching

The exemplars used in both the exemplar matching and in the combined system are half-digits which are extracted from the clean training set and segmented by a conventional HMM-based system. As argued before, the design strives for long units. Full digits turned out to be too long for matching without DTW resulting in a high error rate. With half-digits the exemplars seemed to generalize sufficiently to unseen data resulting in an acceptable baseline (see below). This results in 49,354 exemplars belonging to 22 half-digit classes and 14,418 silence exemplars (in total 63,772 exemplars). The minimum and maximum exemplar lengths are 5 and 30 frames respectively. Exemplars longer than 30 frames are omitted to limit the number of dictionaries that are to be used in the further steps of the experiment.

Speech segments are classified as the their single closest neighboring exemplar (1-NN). The exemplar startup and gender costs are tuned manually for maximal recognition accuracy.

3.3. Combined System

In the combined system, there are in total 508 dictionaries containing the same speech exemplars as in the exemplar matching baseline. However, only 1300 silence exemplars (50 exemplars for each length) are used since silence exemplars do not contribute much as discussed in Section 2.3. The system ends up using 50,654 exemplars in total. In the combined system, the l_2-norm of each dictionary column is set to unity, i.e. the energy of each exemplar is normalized. The same normalization is applied to the reshaped input speech vectors. The reconstruction error shows enough discrimination among classes after 50 iterations. All elements of $\mathbf{\Lambda}$ are set to 2. The scale factor C is set to 0.5. The combined system only uses exemplar startup cost which is, like for the exemplar matching system, tuned for maximal accuracy.

3.4. Reconstruction Error Metrics

The log-compressed features that are used in the exemplar matching baseline and the first intermediate system are compared using the Euclidean distance. All the other intermediate systems and the final system use the generalized KLD to calculate the reconstruction error.

4. RESULTS AND DISCUSSION

In this section, we migrate the 1-NN exemplar matching system in several steps towards the final combined design. The steps dealing with feature representation and distance metric in a single nearest neighbor exemplar matching context are summarized in Table 1.

Based on prior design experience, we start from a design using MIDA features, channel (mean) normalization and Euclidean distance resulting in a word error rate (WER) of 1.11%. Since the sparse representation approach does not use linear transforms or derivatives, we remove this first, resulting in 3.36% WER. The second and the third lines compare the recognition accuracies obtained using mean normalized log-compressed power spectra and mean normalized linear power spectra in conjunction with the Euclidean distance and generalized KLD respectively. It can be concluded that log-compression combined with the Euclidean distance performs much better. The results in the last two blocks of Table 1 shows that the generalized KLD couples much better with linear magnitude spectra and mean normalization is not effective both for magnitude and power spectra in this task.

The first block of Table 2 presents the WER obtained with the combined system using power and magnitude spectra. Compared to the last block of Table 1, there is a significant improvement on recognition accuracies in both magnitude and power spectra due to sparse combination approach. Finally, the l_2-norm of the exemplars is set to unity as described in Section 3.3 which boosts the recognition accuracy. Even though the best result obtained with the proposed approach is still behind the baseline system, it significantly outperforms all the other intermediate systems with comparable features.

5. CONCLUSIONS

We discussed two different exemplar-based recognition schemes, namely exemplar matching and exemplar-based sparse representations, and proposed a combined system that uses multiple length exemplars to jointly approximate the input speech. Such a design can benefit from the noise model provided by the sparse representations approach while it can decode unseen speech directly in terms of exemplar identities using a reconstruction error metric. Exemplars are organized in separate dictionaries which are expected to provide better classification than using a single dictionary as every input segment is approximated by a combination of exemplars belonging to the same class only. The additivity and non-negativity requirement limits the representation domain to magnitude or power spectra. This apparently leads to lower recognition accuracy compared to discriminatively trained speech features. Moreover, the Euclidean distance, which is widely used in exemplar matching based systems, has to be replaced by the generalized KLD.

This initial study still leaves many questions open, some of which are listed here. Firstly, other divergences such as Itakura-Saito have been proposed to model spectral data in exemplar techniques. Secondly, the dictionaries can most likely be pruned without accuracy cost. Thirdly, when dropping non-negativity of the representation, linear transforms of the features are allowed. Fourth, exemplar sequences can be constrained by e.g. prosody.

6. ACKNOWLEDGEMENTS

This work has been supported by the Fund for Scientific Research Flanders, FWO, under grant G.0260.07 (TELEX) and KU Leuven research grant OT/09/028 (VASI).

7. REFERENCES

[1] 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, May 2007.

[2] S. Axelrod and B. Maison, "Combination of hidden markov models with dynamic time warping for speech recognition," in *Proc. ICASSP*, May 2004, vol. 1, pp. 173–176.

[3] G. Aradilla, J. Vepa, and H. Bourlard, "Improving speech recognition using a data-driven approach," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 3333–3336.

[4] K. Demuynck, *Extracting, Modelling and Combining Information in Speech Recognition*, Ph.D. thesis, K.U.Leuven, ESAT, February 2001.

[5] L. Golipour and D. O'Shaughnessy, "Context-independent phoneme recognition using a k-nearest neighbour classification approach," in *Proc. ICASSP*, Apr. 2009, pp. 1341 –1344.

[6] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[7] J. F. Gemmeke, L. ten Bosch, L. Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proc. EUSIPCO*, Glasgow, Scotland, August 2009, pp. 1755–1759.

[8] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representations features for speech recognition," in *Proc. INTERSPEECH*, Sept. 2010, vol. 2, pp. 2254–2257.

[9] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. ICASSP*, March 2010, pp. 4546–4549.

[10] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proc. ICASSP*, May 2011, pp. 4588–4591.

[11] Q. F. Tan and S. S. Narayanan, "Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1337 –1346, May 2012.

[12] K. Laurila, "Noise robust speech recognition with state duration constraints," in *Proc. ICASSP*, Apr. 1997, vol. 2, pp. 871–874.

[13] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

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

[15] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Tutorial and Research Workshop ASR2000*, Sept. 2000, pp. 181–188.