Improvements of a dual-input DBN for noise robust ASR

Yang Sun[1], *Jort F. Gemmeke*[1], *Bert Cranen*[1], *Louis ten Bosch*[1], *Lou Boves*[1]

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
{Y.Sun, J.Gemmeke, B.Cranen, L.tenBosch, L.Boves}@let.ru.nl

Abstract

In previous work we have shown that an ASR system consisting of a dual-input Dynamic Bayesian Network (DBN) which simultaneously observes MFCC acoustic features and an exemplar-based Sparse Classification (SC) phoneme predictor stream can achieve better word recognition accuracies in noise than a system that observes only one input stream. This paper explores three modifications of SC input to further improve the noise robustness of the dual-input DBN system: 1) using state likelihoods instead of phonemes, 2) integrating more contextual information and 3) using a complete set of likelihood distribution. Experiments on AURORA-2 reveal that the combination of the first two approaches significantly improves the recognition results, achieving up to 29% (absolute) accuracy gain at SNR -5 dB. In the dual-input system using the full likelihood vector does not outperform using the best state prediction.

Index Terms: ASR, noise robustness, sparse classification, dual-input DBN

1. Introduction

Systems based on Hidden Markov models (HMMs) that obtain observation likelihoods by modeling speech with Gaussian Mixture Models, have dominated the automatic speech recognition (ASR) field for the last 30 years. While quite successful in dealing with clean, read or prepared speech, the performance of this type of recognizer is known to degrade dramatically under noisy conditions or with spontaneous conversational speech. Despite the many modifications that have been proposed to different modules of HMM-based ASR systems, a large performance gap still remains between ASR and Human Speech Recognition (HSR) [1]. There is growing consensus that besides the likelihoods from the GMMs, complementary information is needed. Many studies, such as [2], have shown that hybrid systems can perform better than either individual system alone. In this work, we aim at early fusion of two systems, one is a conventional GMM-based system and the other one is a phone/state predictor.

In previous work [3], we used the Graphical Modeling Toolkit (GMTK) [4] to implement a Dynamic Bayesian Network (DBN) for noise-robust ASR. Together with the MFCC features modeled by Gaussian Mixture Models (GMM), an additional stream of discrete observations was fed into the DBN. The latter stream contained the index of the most likely phoneme for each MFCC frame, obtained from an exemplar-based, so-called *Sparse Classification* (SC) system [5]. With noisy speech being modeled as a linear combination of both clean speech and noise exemplars, the SC system is inherently noise robust when a suitable dictionary of speech and noise exemplars is used. A DBN operating only on SC inputs (denoted by *SC-only* hereafter), where DBN acts as a Viterbi decoder, outperforms a classical MFCC-based DBN system (*MFCC-only* hereafter) at low SNRs. However, at high SNRs the SC-only system results in lower recognition accuracies.

Since we are interested in early fusion, we use an intermediate representation in the SC system, viz. vectors that for each time frame indicate the likelihood of all candidate states (using the conventional 16-state digit models in the AURORA-2 connected digit recognition task). In [3] we converted the 179-dimensional state likelihood vector to the label of the most likely phoneme. Despite this simplification, the dual-input DBN still successfully combined the strengths of MFCC and SC, resulting in a substantially improved word accuracy at lower SNRs of the dual input system and without losing performance at high SNRs.

In this paper, we propose three methods to further improve the performance of this dual-input system. As in [3], we evaluate the effectiveness of these three methods using the AURORA-2 task. First, we use the index of the most likely *HMM-state*, rather than the label of the most likely *phoneme*. This avoids the potentially ambiguous mappings from the 16 states in the 11 digit models to one of the 20 phonemes (including silence) that describe the digits. Second, in [3] we used exemplars that span 10 frames in the SC system. In [5] it was shown, however, that larger exemplar sizes can lead to a higher noise robustness at low SNRs, be it at the cost of lower accuracies at high SNRs. In this paper we investigate whether the dual-input DBN system also benefits from using larger exemplar sizes at low SNRs without a drop at clean. Third, we use the full 179 dimensional posterior likelihood vector generated by SC as the second input stream instead of the label of the most likely state. We expect, on the one hand, a complete set of likelihoods contains more information besides the winner state in order to improve the recognition especially when the winner is wrong, on the other hand, more interaction between two streams can be achieved by doing so.

The rest of this paper is organized as follows. In Section 2, the dual-input dynamic Bayesian network (DBN) architecture is introduced. It is followed by a short introduction to sparse classification (SC) in Section 3, which provides the second input of our DBN. We describe our experiments and discuss the results in Section 4. Finally a conclusion is drawn in Section 5.

2. Dynamic Bayesian Networks

2.1. DBN architecture

Figure 1 depicts the input stage of the dual-input DBN architecture used in our study. The random variable s_t represents the states over time t and the shaded circular nodes x_t represent the traditional MFCC features modeled by GMMs (*MFCC* hereafter). The shaded square nodes SC_t represent some external evidence, in our case, provided by the SC system (cf. Section 3).

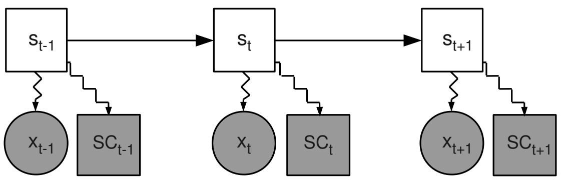


Figure 1: Architecture of the input stage of the dual input DBN.

2.2. Info of the winner vs. the full likelihood distribution

We can inform the DBN the most likely candidate only or we can feed it with all information of SC output by using full SC likelihood vector. In the former case, we use the index of the most likely candidate (phoneme or state) according to the SC likelihood vector for each time frame. While the MFCC node x_t is always modeled by a GMM, a conditional likelihood table (CPT) is used to model the discrete observations of the index.

In order to be able to input the full 179-element likelihood vector in GMTK, we used the technique called Virtual Evidence (VE) (more details about VE can be found in [6]). In this scenario the complete posterior likelihood distribution from the external SC system is used as a prior distribution by the DBN.

Both methods have their own strength. On the one hand, the index provides crisp information which can alleviate the negative impact of uncertainties in the external system. On the other hand, the full likelihood vector offers a complete picture of all the candidates and keeps the possibility for correcting the mistakes made by the SC system in providing the index of the most likely phoneme or state.

3. Sparse Classification

In the Sparse Classification (SC) system [5], an observed speech segment is expressed as a sparse, non-negative linear combination of segments of speech, named *exemplars*, which are extracted from a training database. Each exemplar spans multiple frames. In this work, we compared two exemplar sizes, $T = 10$ and $T = 30$ frames, denoted by *T10* and *T30* respectively hereafter. Likewise, segments of noise are modelled as a linear combination of noise exemplars. Using a collection of noise and speech exemplars, called a *dictionary*, we express noisy speech as a linear combination of both speech and noise exemplars. By finding the sparsest possible set of speech and noise exemplars that approximates the observed noisy speech, we obtain a sparse representation of each observed speech segment.

Each exemplar in the speech part of the dictionary is labelled with HMM-state labels obtained from a conventional MFCC-based decoder. Using the recovered sparse representation, we use the weights of each exemplar to obtain posterior state likelihoods by calculating the weighted linear combination of underlying state labels.

4. Experiments and Results

4.1. Database and feature extraction

Both MFCC and SC inputs used in training were obtained from the clean training set of the AURORA-2 corpus (8440 utterances). For testing purposes, we only used test set 'A', i.e., utterances of four noise types (subway, car, babble, exhibition hall) at SNR levels -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB and $\inf$ (clean speech). Each subset contains 1001 utterances consisting of a sequence of one to seven digits, 'zero-nine' and 'oh'.

The MFCC input to the DBN consists of 39 dimensional vectors containing 12 MFCC features plus a separate log-energy coefficient as well as the corresponding first and second order delta coefficients. They are based on a 23 band Mel frequency spectrum using a frame shift of 10ms and a frame length of 25ms. Subsequently, the MFCCs are normalized with respect to their mean and variance per utterance. The MFCC feature vectors are represented by diagonal covariance Gaussian Mixtures. Our final model consists of up to 32 diagonal covariance Gaussian Mixtures.

To obtain the SC information, we used the same configuration as in [5]. In a nutshell, the SC method operates on 23-dimensional Mel-scale magnitude features, and uses a dictionary created with 4000 noise and 4000 speech exemplars randomly extracted from the multi-condition training set with exemplar sizes *T10* or *T30*. The output of the SC system is a 179 dimensional vector for each frame, corresponding to the likelihood of each HMM-state.

Instead of just reporting the results of the dual-input system (denoted by *MFCC/SC* hereafter), we also provide results of each single input (SC-only and MFCC-only) system separately for comparison. The word recognition accuracies are averaged over the four noise types in all results.

4.2. Experiment 1: From most likely phoneme to state

Two issues motivates us moving from phoneme input to state one: 1) in DBN, we use states instead of phonemes to compose each digit. It will be more proper if we feed the information about the target layer itself. 2) phoneme input loses the relationship between words and states because some digits comprise the same phonemes (e.d. /s/ in 'six' and 'seven'). Therefore, rather than using the index of the most likely *phoneme* as in [3], in this experiment we use the index of the most likely *state*. Hence, the cardinality of the SC input variable is increased from 20 to 179.

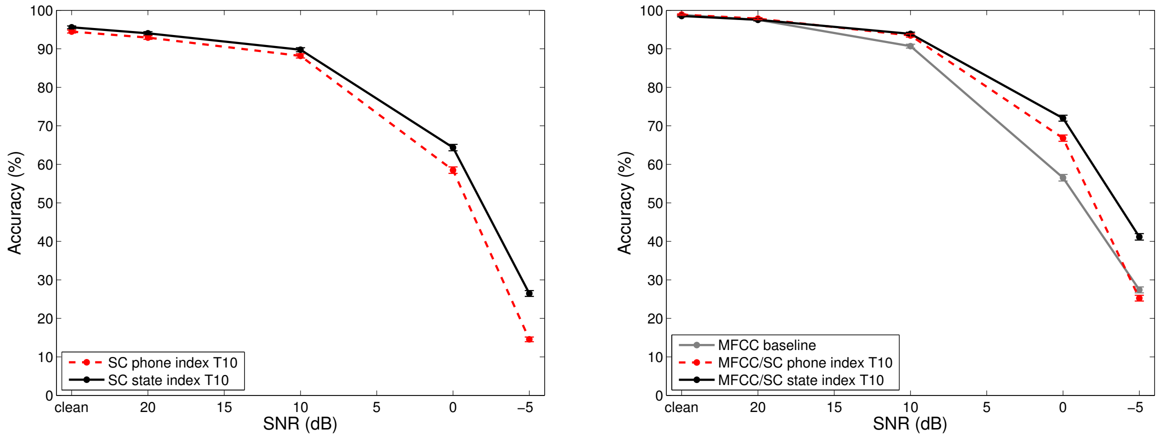
Figure 2 shows the word recognition accuracy for both methods. The use of the most likely state index substantially outperforms the use of the most likely phoneme index. This is most probably due to the fact that we modelled the digits by means of 16 state word models. Thus the index of the most likely state provided by the SC system informs the DBN more directly about the digit's a priori likelihood than a most likely phoneme index.

Using the state index especially improved the accuracy at lower SNRs, up to 12% and 16% (absolute) at SNR -5 dB, both for the SC-only and the combined MFCC/SC system. The grey curve in Figure 2b (also shown in Figures 3b and 4b) indicates the baseline system which only uses MFCC observations.

It is clear that the dual-input system with the state index performs much better than the baseline, but it also outperforms the dual-input with the phoneme index, especially at low SNRs. Therefore, we will use the state index instead of the phoneme index in the following experiments.

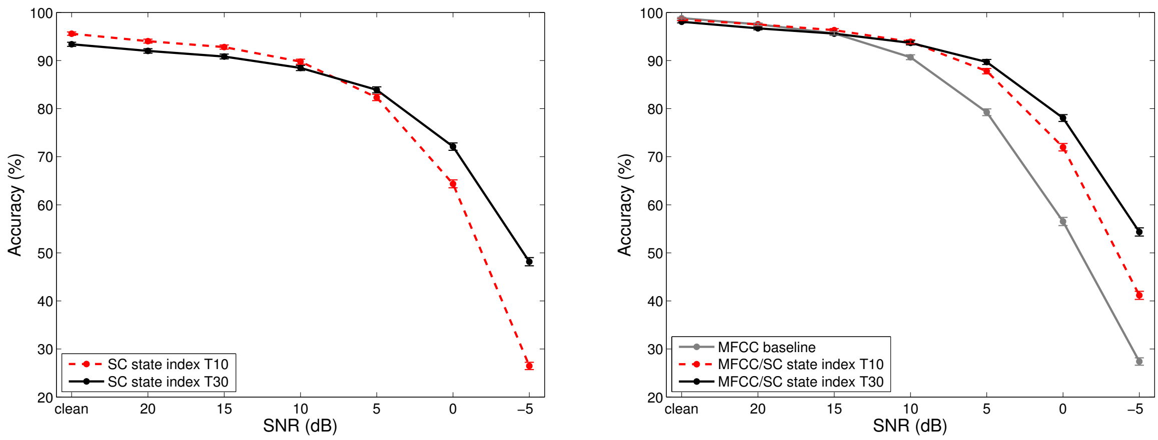
4.3. Experiment 2: Increasing the amount of contextual information

In [3] it was shown that the dual-input DBN that combines MFCCs and *T10* SC can retain the MFCC's good performance at high SNR as well as retaining the good performance at low SNRs of *T10* SC. In another words, the strengths of two worlds are combined. In [5] the influence of the exemplar size was in-



(a) SC-only (b) MFCC/SC

Figure 2: Word accuracy on AURORA-2 as a function of SNR. The performance when using the index of the most likely SC state is shown by a solid line while performance obtained with the index of the most likely SC phoneme is shown by a dashed line. Figure (a) depicts the performance of the SC-only system and Figure (b) pertains the dual-input DBN. In Figure (b), the performance of the MFCC-only baseline is shown by a solid gray line.



(a) SC-only (b) MFCC/SC

Figure 3: Word accuracy on AURORA-2 as a function of SNR. The performance when using the *T30* SC exemplars is shown by a solid line while performance of *T10* SC exemplars is shown by a dashed line. Figure (a) depicts the performance of the SC-only system and Figure (b) pertains the dual-input DBN. In Figure (b), the performance of the MFCC-only baseline is shown by a solid gray line.

vestigated. It was found that *T10* SC segments performed better at high SNRs, whereas longer *T30* segments were optimal at lower SNRs. In this section, we want to investigate the effect of the of input *T30*, which contains longer temporal context, in the dual-input system.

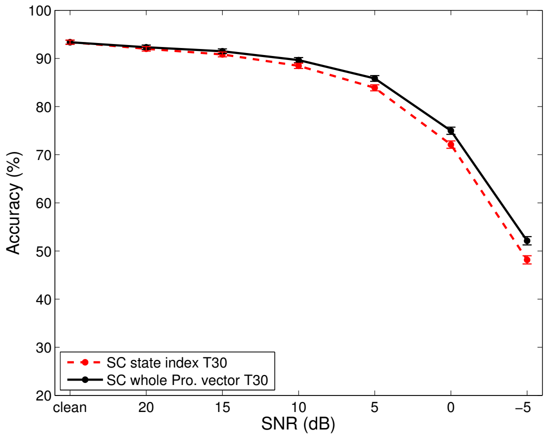
Figure 3a shows the comparison between SC-only system with exemplar sizes of T10 and T30. As in [5], the results show that the performance curves cross at approximately SNR 10 dB. For higher SNRs, the exemplar size *T30* performs slightly worse than *T10*. However for lower SNRs, *T30* works much better, e.g. *T30* outperforms *T10* by around 22% at SNR -5dB.

For the combined MFCC/SC system, the results are shown in Figure 3b. Now *T30* performs equally well as *T10* in the cleaner conditions, meaning that the decrease of accuracy at high SNRs has been compensated by the simultaneous use of the MFCC observation stream. On the other hand, *T30* still outperforms *T10* significantly in very noisy conditions. For instance, the improvement is 13% (absolute) at SNR -5 dB. Consequently, exemplars with a size of *T30* will be used instead of *T10* in the next experiment.

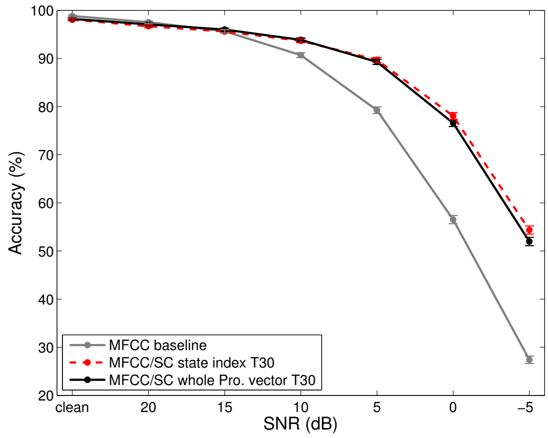
It is also worth mentioning that the dual-input system outperforms each stand-alone system in most noise conditions, especially at SNR -5 dB. This demonstrates the benefit of an integrated system over a switching system that can choose among the outputs of several different systems that operate in parallel, but can never perform beyond the best one.

4.4. Experiment 3: From index to full likelihood vector

So far, the additional input to the DBN consisted of the index of the phoneme or state that was most likely according to the SC system. This means we schrink SC likelihood vector from 179 dimensions into 1, all of our trust is laid on the prediction made by SC system and all the rest 178 dimensional states are negelected. However, SC index is incorrect for many frames, espe-



(a) SC-only



(b) MFCC/SC

Figure 4: (a) Word accuracy on AURORA-2 as a function of SNR. The performance when using the whole SC likelihood vector is shown by a solid line while performance obtained with the index of the most likely SC state is shown by a dashed line. Figure (a) depicts the performance of the SC-only system and Figure (b) pertains the dual-input DBN. In Figure (b), the performance of the MFCC-only baseline is shown by a solid gray line.

cially for low SNR cases. Therefore, in a third experiment, we investigate whether the use of the full likelihood vector will improve the recognition accuracy through recovering the runner-ups possibilities.

The performance of the SC-only system is shown in Figure 4a. In high SNR conditions there is no statistically significant difference between using the index of the winning state and the full likelihood vector. However, as the energy of the noise increases, the advantage of the full likelihood vector become obvious; for example, 4% (absolute) accuracy is gained at SNR -5 dB.

The performance of the MFCC/SC system is shown in Figure 4b. We still get the best of two worlds. However, the benefit of using the full likelihood vector over the best state index in the -5 dB SNR condition that we observed in Figure 4a is no longer present. In fact, the index-based system even works slightly better at SNRs -5 and 0 dB. One possible explanation for this is that, in the combined system, more provided SC dimensions also enlarge the searching space in MFCC-based GMM which has been proved to be irreliable in low SNRs. Thus, meanwhile we involve more information from the SC side, we also involve more confusion from MFCC side, which clearly brings more cons over the pros we added. But some previlege experiments have shown that adding 2 best dimensions of SC likelihoods outperforms the combined system with SC-index by 3% and 0.5% in SNR -5dB and clean speech respectively. We can conclude using SC index in the combined system only gives us a suboptimal performance but full SC likelihood vector also brings too much fuzziness into the competition. There should be one maximal performance exist in between.

5. Conclusions

In this work, we proposed three methods to further improve the noise robustness of the ASR system described in [3], which consists of a dual-input DBN simultaneously observing MFCC acoustic features and an exemplar-based Sparse Classification (SC) phoneme predictor stream. Experiments on AURORA-2 reveal that the combination of the first two approaches significantly improves the recognition results, achieving up to 29% (absolute) accuracy gain at SNR -5 dB without any degradations at high SNRs. When used as the only input, the full vector of SC state probabilities outperforms the index of the winning state, whereas in the dual-input system the benefit of the full vector over the winning state index is no longer present.

As an extension of this work, it is necessary to imvestigate the optimal dimension of SC in the combined system for each SNR level. Moreover, although quite successful in improving the noise robustness of the system, the way in which we combined the two input streams in the current study does not allow the DBN to learn the dependency relations between the two streams. In future work, we will investigate to what extent explicitly modelling such dependency relations may help to further improve recognition performance.

6. Acknowledgements

Yang Sun has received funding from European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 213850 - SCALE. Jort F. Gemmeke is supported by the Dutch-Flemish STEVIN Program.

7. References

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, no. 1, pp. 1–15, 1997.

[2] H. Bourlard and N. Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," in *International School on Neural Nets: Adaptive Processing of Temporal Information*. Springer Verlag, 1997.

[3] Y. Sun, J. F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, "Using a DBN to integrate sparse classification and gmm-based ASR," in *Proceedings of Interspeech 2010*, Makuhari, Japan, 2010.

[4] J. Bilmes, "The GMTK documentation," http://ssli.ee.washington.edu/~bilmes/gmtk.

[5] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proceedings of ICASSP 2010*, Dallas, USA, 2010.

[6] J. Bilmes, "On soft evidence in bayesian networks," University of Washington, Dept. of Electrical Engineering, Tech. Rep. UWEETR-2004-0016, 2004.