Improvements of a dual-input DBN for noise robust ASR

Yang Sun\textsuperscript{1}, Jort F. Gemmeke\textsuperscript{1}, Bert Cranen\textsuperscript{1}, Louis ten Bosch\textsuperscript{1}, Lou Boves\textsuperscript{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 recognition 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) \cite{Boves2008}. There is growing consensus that besides the likelihoods from the GMMs, complementary information besides the winner state in order to improve the performance at high SNRs.

In previous work \cite{Sun2012}, we used the Graphical Modeling Toolkit (GMTK) \cite{Barnett2007} 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 \cite{Gemmeke2010}. 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 (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 \cite{Sun2012} 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 \cite{Sun2012}, 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 \cite{Sun2012} we used exemplars that span 10 frames in the SC system. In \cite{Sun2011} 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).
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.g. /si/ 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-
investigated. 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 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 -5 dB.

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 shrink 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 neglected. However, SC index is incorrect for many frames, espe-
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 signif-
ically improves the recognition results, achieving up to 29%
(absolute) accuracy gain at SNR -5 dB without any degrada-
tions 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 investi-
gate the optimal dimension of SC in the combined system for
each SNR level. Moreover, although quite successful in impro-
ing 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 ex-
tent explicitly modelling such dependency relations may help to
further improve recognition performance.

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 signif-
ically improves the recognition results, achieving up to 29%

(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.