INTERSPEECH 2009 BRIGHTON

Robust In-Car Spelling Recognition - A Tandem BLSTM-HMM Approach

Martin Wöllmer[1], *Florian Eyben*[1], *Björn Schuller*[1], *Yang Sun*[1],
Tobias Moosmayr[2], *Nhu Nguyen-Thien*[3]

[1]Technische Universität München, Institute for Human-Machine Communication,
80290 München, Germany
[2]BMW Group, Forschungs- und Innovationszentrum, 80788 München, Germany
[3]Continental Automotive GmbH, Interior BU Infotainment & Connectivity,
Advanced Development and Innovation, 93055 Regensburg, Germany

woellmer@tum.de

10.21437/Interspeech.2009-375

Abstract

As an intuitive hands-free input modality automatic spelling recognition is especially useful for in-car human-machine interfaces. However, for today's speech recognition engines it is extremely challenging to cope with similar sounding spelling speech sequences in the presence of noises such as the driving noise inside a car. Thus, we propose a novel Tandem spelling recogniser, combining a Hidden Markov Model (HMM) with a discriminatively trained bidirectional Long Short-Term Memory (BLSTM) recurrent neural net. The BLSTM network captures long-range temporal dependencies to learn the properties of in-car noise, which makes the Tandem BLSTM-HMM robust with respect to speech signal disturbances at extremely low signal-to-noise ratios and mismatches between training and test noise conditions. Experiments considering various driving conditions reveal that our Tandem recogniser outperforms a conventional HMM by up to 33%.

Index Terms: spelling recognition, recurrent neural networks, long short-term memory, noise robustness

1. Introduction

In many voice command applications speech input cannot be restricted to a fixed set of words. For example in-car internet browsers which are already available in today's upper class cars, demand for fast, intuitive, and optionally hands-free operation. While basic browser commands may be covered by a few keywords, entering a URL via speech input presumes an Automatic Speech Recognition (ASR) system that also allows spelling. However, since many letters such as "b" and "d" sound fairly similar, spelling recognition in the presence of driving noise is very challenging - even for humans. In contrast to natural speech, spelling recognition cannot be improved by the usage of a language model but exclusively relies on discriminating the acoustic patterns of different letter utterances. Only for simplified cases such as matching the spelled sequence against a stored dictionary [1] "language information" can be used.

In order to make ASR systems applicable in noisy environments a large number of different techniques to improve noise robustness has been proposed in recent years. An overview of speech feature enhancement approaches that can potentially cope with comparably stationary noise sources like driving noise can be found in [2]. While most strategies for noise compensation, like the Switching Linear Dynamic Model proposed in [3], show good performance for isolated noisy speech utterances with predefined speech on- and offset, their real-life applicability suffers from the lack of a reliable discrimination between speech and noise segments [4]. Especially in the interior of a car where signal-to-noise ratio (SNR) levels are typically negative, proper voice activity detection is a non-trivial task.

Apart from feature enhancement techniques, also alternative model architectures have been investigated to make ASR more robust. Tandem architectures which combine the output of a discriminatively trained neural net with dynamic classifiers such as Hidden Markov Models (HMMs) have been successfully used for speech recognition tasks and are getting more and more popular [5, 6]. Yet, the amount of contextual information a conventional recurrent neural network (RNN) can incorporate into the HMM decoder in order to learn noise dynamics and to improve phoneme discrimination is limited. The major reason for this is that the backpropagated error in RNNs either blows up or decays over time. Long Short-Term Memory (LSTM) recurrent neural nets [7] overcome this problem by using memory cells to store and access information over long time periods. Thus, this paper introduces a novel Tandem decoder which combines bidirectional Long Short-Term Memory (BLSTM) recurrent neural nets and HMMs for noise robust spelling recognition. Thereby the modelling of long-range context information is used to learn typical in-car noise characteristics, allowing a better discrimination between speech and noise in the time and frequency domain. Our Tandem recogniser uses the phoneme predictions of a BLSTM net together with conventional MFCC features to reliably detect spelled letter sequences in driving noise. Thereby our technique can not only cope with extremely low SNR levels but also with a mismatch between noise conditions during training and testing.

The paper is organised as follows: Section 2 outlines the principle of BLSTM networks, Section 3 introduces the Tandem BLSTM-HMM which is used in the in-car spelling recognition experiments in Section 4 before a conclusion is drawn in Section 5.

2. Bidirectional LSTM Networks

Bidirectional recurrent neural networks [8] are composed of two recurrent network layers, whereas the first one processes the sequence forwards and the second one processes it backwards. Since both networks are connected to the same output layer, the bidirectional net has access to the entire information about past and future sequence data points. During training, the amount of contextual information that the network uses is learnt and does not have to be specified manually. Bidirectional networks can

be applied whenever the sequence processing task is not truly online (meaning the output is not required after every input) which makes them popular for speech recognition tasks where the output has to be present e.g. at the end of a sentence [9].

A drawback of conventional bidirectional RNN architectures is that the range of context that can actually be accessed is limited as the influence of a given input on the hidden layer either decays or blows up exponentially over time (*vanishing gradient problem*). This led to the introduction of Long Short-Term Memory (LSTM) RNNs [7]. An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more recurrently connected memory cells, along with three multiplicative "gate" units: the input, output, and forget gates. Figure 1 shows the architecture of a simple LSTM memory block consisting of a single memory cell. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate. Their effect is to allow the network to store and retrieve information over long periods of time. If, for example the input gate remains closed, the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate. This principle solves the vanishing gradient problem and gives access to long range context information.

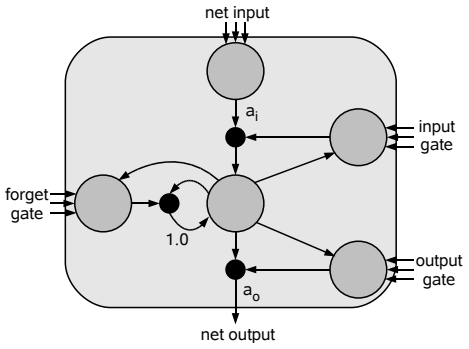

Figure 1: *LSTM memory block consisting of one memory cell: input, output, and forget gate collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state respectively;* a_i *and* a_o *denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state*

The combination of bidirectional networks and LSTM is called bidirectional LSTM, which has demonstrated excellent performance in phoneme recognition [9], keyword spotting [10], and emotion recognition [11]. A detailed explanation of BLSTM networks can be found in [12].

3. Tandem BLSTM-HMM Decoder

The Tandem BLSTM-HMM decoder applied in this work is depicted in Figure 2. The lower, grey-shaded part of the figure shows the basic LSTM architecture consisting of an input i_t, an output o_t, and a hidden node h_t for each time step. For the sake of simplicity only a simple LSTM layer is illustrated in Figure 2, instead of the more complex bidirectional LSTM which would be composed of two RNNs as explained in Section 2. The upper part of Figure 2 shows the explicit Dynamic Bayesian Network (DBN) representation of the Hidden Markov Model. In contrast to the *implicit* HMM graph representation which uses a single Markov chain together with an integer state to represent all information, the *explicit* approach [13] models information such as the current word, the indication of a word transition, or the position within a word by hidden random variables.

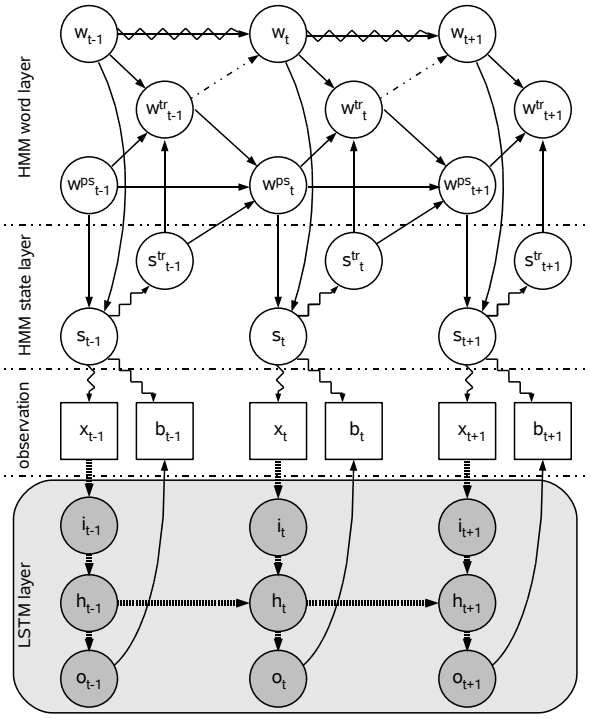

Figure 2: *Architecture of the Tandem BLSTM-HMM*

For every time step, the following random variables are defined: w_t represents the current word, w_t^{ps} denotes the position within the word, w_t^{tr} is a binary indicator variable for a word transition, and s_t is the HMM state with s_t^{tr} indicating a state transition. The variable x_t denotes the observed acoustic features and b_t contains the phoneme prediction of the BLSTM which is used as an additional discrete observation. The DBN structure in Figure 2 displays hidden variables as circles and observed variables as squares. Straight lines represent deterministic conditional probability functions (CPFs) whereas random CPFs correspond to zig-zagged lines. Dotted lines refer to so-called *switching parents* which in our case switch between two different CPFs. Note that the bold dashed lines in the LSTM layer of Figure 2 do not represent statistical relations but simple data streams.

For a speech sequence of length T, the DBN structure expresses the following factorisation:

$$\begin{aligned} &p(w_{1:T}, w_{1:T}^{tr}, w_{1:T}^{ps}, s_{1:T}^{tr}, s_{1:T}, x_{1:T}, b_{1:T}) = \\ &\prod_{t=1}^{T} p(x_t|s_t)p(b_t|s_t)f(s_t|w_t^{ps}, w_t)f(w_t^{tr}|w_t^{ps}, w_t, s_t^{tr}) \\ &p(s_t^{tr}|s_t)f(w_1^{ps})p(w_1)\prod_{t=2}^{T} p(w_t|w_{t-1}^{tr}, w_{t-1}) \\ &f(w_t^{ps}|s_{t-1}^{tr}, w_{t-1}^{ps}, w_{t-1}^{tr}) \end{aligned} \tag{1}$$

Thereby $p(\cdot)$ describes random conditional probability

functions and $f(\cdot)$ denotes deterministic CPFs.

The size of the BLSTM input layer i_t corresponds to the dimensionality of the acoustic feature vector x_t whereas the vector o_t contains one probability score for each of the P different phonemes at each time step. b_t is the index of the most likely phoneme:

$$b_t = \max_{o_t}(o_{t,1}, ..., o_{t,j}, ..., o_{t,P}) \quad (2)$$

The CPFs $p(x_t|s_t)$ are described by Gaussian mixtures as common in an HMM system. Together with $p(b_t|s_t)$ and $p(s_t^{tr}|s_t)$, they are learnt via EM training.

The binary variable s_t^{tr} is equal to one whenever there is a state transition and zero otherwise. A simple deterministic CPF $f(s_t|w_t^{ps}, w_t)$ maps from a given position in a word w_t to the corresponding whole word state. Similarly, the word position can be inferred deterministically via $f(w_t^{ps}|s_{t-1}^{tr}, w_{t-1}^{ps}, w_{t-1}^{tr})$. A word transition occurs whenever $s_t^{tr} = 1$ and $w_t^{ps} = S$ provided that S denotes the number of states of a word. w_{t-1}^{tr} is a switching parent of w_t, meaning that if no word transition occurs, w_t is equal to w_{t-1}. Otherwise a word bigram which makes each word equally likely, but assumes a short silence between two words (or letters in case of the spelling recognition experiment) is used.

4. Experiments

For the evaluation of the noise robustness of our Tandem BLSTM-HMM spelling recogniser we used the letter utterances from "a" to "z" from the TI 46 Speaker Dependent Isolated Word Corpus to generate a large set of spelling sequences. The database contains utterances from 16 different speakers - eight females and eight males. Per speaker 26 utterances were recorded for every letter whereas ten samples are used for training and 16 for testing. Consequently the overall isolated letter training corpus consists of 4 160 utterances while the test set contains 6 656 samples. In order to obtain connected spelling sequences, the isolated letters from every speaker were randomly combined to sequences including between three and seven letters. The silence at the beginning and at the end of the isolated letters was not cut, leading to short silence segments in between the letters. Thereby each individual letter utterance occurs only once within the whole corpus of connected letters. The resulting corpus consists of 839 sequences for training and 1 354 for testing. Note that our database bears some similarity to the *Aurora* task of recognising connected TI digits utterances in noise. Yet, spelling recognition is more challenging since letters are harder to discriminate than digits.

Out of the clean spelling utterances, noisy sequences were generated by superposing the speech signal with different in-car noise types. In general, interior noise can be split up into four groups: the first one is wind noise which is generated by air turbulences at the corners and edges of the vehicle. A further noise type is engine noise which depends on load and number of revolutions. The third noise group is caused by wheels, driving, and suspension and is influenced by road surface and wheel type. Thus, a rough surface causes more wheel and suspension noise than a smooth one. Finally, buzz, squeak and rattles generated by pounding or relative movement of interior components of a vehicle have to be considered.

According to existing in-car speech recognition systems, the microphone was placed in the middle of the instrument panel. Note that the mouth-to-microphone transfer function has been neglected, since the masking effect of background noise was proven to be much higher than the effect of convolutional noise. In an additional experiment the slight degradation of recognition performance in case of a convolution of the speech signal with a recorded in-car impulse response could be perfectly compensated by simple cepstral mean subtraction.

As interior noise masking varies depending on vehicle class and derivates, the speech sequences were superposed by noise of four different BMW vehicles: a 5 series touring, a 6 series convertible, an M5 sedan, and a MINI Cooper convertible. Recognition results are presented exclusively for the worst case in-car noise scenario which was proven to occur in the MINI convertible.

Surface	Velocity	Abbreviation
Big cobbles	30 km/h	COB
Smooth city road	50 km/h	CTY
Highway	120 km/h	HWY

Table 1: *Considered road surfaces and velocities*

The road surface has an even stronger influence on the characteristics of interior noise. Hence, three different surfaces in combination with typical velocities have been considered as shown in Table 1. The lowest excitation provides a drive over a smooth city road at 50 km/h and medium revolution (CTY). The subsequent higher excitation is measured for a highway drive at 120 km/h (HWY). The worst and loudest sound in the interior of a car is provoked by a road with big cobbles (COB). At 30 km/h wind noise can be neglected, but the rough cobble surface involves dominant wheel and suspension noise. Figure 3 shows the SNR histograms of the noisy speech utterances for each driving condition.

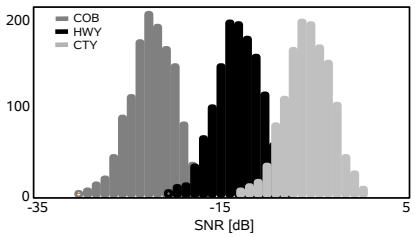

Figure 3: *SNR level histograms for the noisy speech utterances*

In spite of SNR levels below 0 dB, speech in the noisy test sequences is still audible since the recorded noise samples are lowpass signals with most of their energy in the frequency band from 0 to 500 Hz. Consequently, there is no full overlap of the spectrum of speech and noise.

From the speech signal 12 cepstral mean normalised MFCC features together with energy as well as first and second order delta coefficients were extracted. Thereby best results could be obtained when applying a simple FIR highpass filter with a cutoff frequency of 200 Hz in order to partly remove frequency bands that correspond to motor drone etc. before extracting the acoustic features. Without filtering, performance was shown to be significantly lower. However, filtering was only conducted prior to the extraction of the feature vectors x_t processed by the HMM layer of the Tandem recogniser, whereas the BLSTM network processed MFCC features from unfiltered speech before providing the phoneme prediction b_t as additional feature for the HMM layer.

Each letter HMM consisted of eight states while silence was modeled with three states. In addition to the "clean" model,

one BLSTM-HMM was trained for every noise condition using the corresponding noisy training material. During training, all Gaussian mixtures were split once 0.02% convergence was reached. The final models consisted of up to 32 Gaussian mixtures, depending on which models performed best for which training noise condition.

The BLSTM input layer had a size of 39 (one input for each acoustic feature) and the size of the output layer was 25, corresponding to the 25 different phonemes occurring in the spelled letters from "a" to "z". Thereby the network was trained on the forced aligned framewise phoneme transcriptions of the spelling sequences. Both hidden LSTM layers contained 100 memory blocks of one cell each. To improve generalisation, zero mean Gaussian noise with standard deviation 0.6 was added to the inputs during training. We used a learning rate of 10^{-5} and a momentum of 0.9. The independently trained BLSTM network was then incorporated into the Tandem recogniser in order to allow a joint training of the Gaussian mixtures $p(x_t|s_t)$ and the CPFs $p(b_t|s_t)$. To avoid Viterbi paths with zero probability, the CPF $p(b_t|s_t)$ was floored to 10^{-5}.

model	test cond.	HMM	BLSTM-HMM
clean	clean	98.19%	98.80%
CTY	CTY	92.64%	96.55%
HWY	HWY	84.06%	91.15%
COB	COB	81.65%	91.96%
CTY	HWY	60.50%	77.13%
CTY	COB	64.38%	79.70%
HWY	CTY	54.25%	87.51%
HWY	COB	59.09%	85.44%
COB	CTY	79.07%	90.34%
COB	HWY	74.32%	87.58%
	mean	**74.82%**	**88.62%**

Table 2: *Spelling recognition accuracies for the Tandem BLSTM-HMM and the HMM (matched and mismatched condition) - results for clean models in noisy test conditions are not included because of inadequate silence modelling which would lead to permanent insertion errors*

Table 2 shows the word accuracies for the Tandem BLSTM-HMM recogniser and the corresponding HMM without a BLSTM layer. The first column reveals the noise type during training and the second column contains the noise condition during testing. The upper half of the table indicates the "matched condition" case which is valid whenever the recognition system has exact information about the current velocity and road surface. Determining the current velocity is trivial whereas information about the road surface is harder to obtain. Exploiting GPS information might be a possible approach. The lower half of the table shows the "mismatched condition" case when noise types during training and testing are different. Note that a model trained on perfectly clean data fails in noisy test conditions since the silence model will tolerate no signal variance at all, which would lead to permanent insertion errors. In clean conditions both recogniser architectures show almost perfect performance. As soon as the speech signal is corrupted by noise, performance decreases whereas in the matched condition case the BLSTM-HMM outperforms the HMM by up to 10%. Also for the mismatched condition case, the Tandem recogniser is far more robust with respect to noise than the HMM. The greatest improvement can be observed for a recogniser trained on the HWY noise type and tested on a smooth inner city road (CTY). There, the Tandem architecture can increase accuracy by 33%. Conducting the McNemar's test revealed the performance improvement in all experiments is statistically significant at a significance level of 10^{-4}.

5. Conclusion

We introduced a novel Tandem spelling recogniser which combines an HMM architecture with a discriminatively trained bidirectional Long Short-Term Memory recurrent neural net and optimised it for in-car usage. Due to the ability of a BLSTM network to capture long-range temporal dependencies, noise characteristics can be learned and spoken utterances can be detected and discriminated even at negative SNR levels. Our approach tolerates a certain mismatch between training and test noise conditions and outperforms a conventional HMM.

Future works might include the evaluation of the BLSTM-HMM for other noise types or for large vocabulary ASR. Further, it will be interesting to investigate the performance of a Tandem recogniser that processes the entire vector of BLSTM output activations instead of exclusively using the most likely phoneme prediction.

6. References

[1] C. D. Mitchell and A. R. Setlur, "Improving spelling recognition using a tree-based fast lexical match," in *Proc. of ICASSP*, vol. 2, Phoenix, AZ, USA, 2008, pp. 597–600.

[2] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: A comparative survey of robust model architectures and feature enhancement," *Journal on Audio, Speech, and Music Processing (to appear)*, 2009, id 942617.

[3] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *Proc. of ICASSP*, Montreal, Canada, 2004.

[4] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Speech recognition in noisy environments using a switching linear dynamic model for feature enhancement," in *Proc. of Interspeech*, Brisbane, Australia, 2008, pp. 1789–1792.

[5] H. Ketabdar and H. Bourlard, "Enhanced phone posteriors for improving speech recognition systems," in *IDIAP-RR*, no. 39, 2008.

[6] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. of ICASSP*, vol. 3, Istanbul, Turkey, 2000, pp. 1635–1638.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9(8), pp. 1735–1780, 1997.

[8] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, 1997.

[9] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *Proc. of ICANN*, Warsaw, Poland, 2005, pp. 602–610.

[10] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[11] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. of Interspeech*, Brisbane, Australia, 2008, pp. 597–600.

[12] A. Graves, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, Technische Universität München, 2008.

[13] J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, 2005.