Robust Multi-stream Keyword and Non-linguistic Vocalization Detection for Computationally Intelligent Virtual Agents

Martin Wöllmer[1], Erik Marchi[2], Stefano Squartini[2], and Björn Schuller[1]

[1] Institute for Human-Machine Communication,
Technische Universität München, 80333 München, Germany
[2] 3MediaLabs - A3LAB, DIBET - Dipartimento di Ingegneria Biomedica,
Elettronica e Telecomunicazioni, Università Politecnica delle Marche,
60131 Ancona, Italy
{woellmer,schuller}@tum.de, s.squartini@univpm.it

Abstract. Systems for keyword and non-linguistic vocalization detection in conversational agent applications need to be robust with respect to background noise and different speaking styles. Focussing on the Sensitive Artificial Listener (SAL) scenario which involves spontaneous, emotionally colored speech, this paper proposes a multi-stream model that applies the principle of Long Short-Term Memory to generate context-sensitive phoneme predictions which can be used for keyword detection. Further, we investigate the incorporation of noisy training material in order to create noise robust acoustic models. We show that both strategies can improve recognition performance when evaluated on spontaneous human-machine conversations as contained in the SEMAINE database.

Keywords: Conversational agents, keyword spotting, multi-condition training, long short-term memory.

1 Introduction

Systems for advanced Human-Machine Interaction which offer natural and intuitive input and output modalities require robust and efficient machine learning techniques in order to enable spontaneous conversations with a human user. Since speech is the most natural human-to-human communication channel, the advancement of speech technology is an essential precondition for improving Human-Machine Interaction. Conversational agents which shall recognize, interpret, and react to human speech rely on speech processing technologies that can cope with various challenging conditions, such as background noise, disfluencies, and emotional coloring of speech. Reliably extracting meaningful keywords tends to be the most important functionality of speech processing modules providing linguistic information to the dialogue management [1].

As conversational agents are often used in noisy conditions, automatic speech recognition (ASR) and keyword spotting systems have to be based on features and models that lead to an acceptable recognition performance even if the speech

D. Liu et al. (Eds.): ISNN 2011, Part II, LNCS 6676, pp. 496–505, 2011.

signal is superposed by background noise. Thus, most systems apply speech feature normalization or enhancement techniques such as cepstral mean normalization, histogram equalization, or Switching Linear Dynamic Models [2]. A simple and efficient method to improve the noise robustness of the speech recognition back-end is to use matched or multi-condition training strategies [3] by incorporating noisy training material which reflects the noise conditions expected while running the system.

Another approach to enhance recognition performance in challenging conditions is to apply neural networks for generating state posteriors or phoneme predictions which are then decoded by a Hidden Markov Model (HMM). These so-called Tandem or hybrid systems are a popular alternative to the conventional HMM technique since they efficiently combine the advantages of both, neural networks and HMMs [4]. However, conventional Multilayer Perceptrons (MLP) or recurrent neural networks (RNN) as they are used in today's Tandem ASR systems have some inherent drawbacks such as the *vanishing gradient problem* [5] which limits the amount of contextual information that can be modeled by an RNN. Yet, due to co-articulation effects in human speech, context modeling is essential for accurate phoneme prediction. As an alternative to learning a fixed amount of context by processing a predefined number of consecutive feature frames via MLPs, the usage of Long Short-Term Memory (LSTM) networks [6] has recently been proposed for keyword spotting [7] and continuous ASR systems [8]. LSTM networks are able to model a self-learned amount of context information which leads to higher phoneme recognition accuracies when compared to standard RNNs [8].

In this contribution we investigate both, multi-condition training strategies for enhanced keyword spotting performance in noisy conditions, and the effect of incorporating LSTM phoneme prediction in a multi-stream ASR framework. Both techniques are evaluated with respect to their suitability for conversational agents. Thereby we focus on the *Sensitive Artificial Listener* (SAL) scenario which aims at maintaining a natural conversation with different virtual characters [9].

Section 2 describes the four virtual SAL characters that allow for emotional human-machine conversations via the SEMAINE system[1]. For our keyword spotting experiments we use spontaneous speech as contained in the SEMAINE database which is introduced in Section 3 and provides training material for the SEMAINE system. The multi-stream LSTM-HMM technique used for enhanced keyword and non-linguistic vocalization detection within the SEMAINE system is outlined in Section 4. Finally, Section 5 contains the results of our multi-condition training and multi-stream decoding experiments.

2 Sensitive Artificial Listeners

In contrast to most task-oriented dialogue systems, the *Sensitive Artificial Listeners* representing the SEMAINE system [9] focus on aspects of communication that are emotion-related and non-verbal. The system is designed for a one-to-one dialogue situation in which one user is conversing with one of four available

[1] http://semaine-project.eu/

virtual agent characters. Besides speech, the (multimodal) interaction involves head movements and facial expressions. The SAL characters have to recognize a limited set of emotionally relevant keywords, non-linguistic vocalizations such as *laughing* or *sighing*, and the prosody with which the words are spoken. Based on the interpreted input from audio and video, the system has to show appropriate listener behavior, e. g., multimodal *backchannels*, decide when to *take the turn*, and select a suitable phrase in order to maintain the conversation.

The four SAL characters roughly represent areas in the *arousal-valence* space: 'Spike' is angry (high arousal, low valence), 'Poppy' is happy (high arousal, high valence), 'Obadiah' is sad (low arousal, low valence), and 'Prudence' is matter-of-fact (moderate arousal, moderate valence). During the conversations, the virtual characters aim to induce an emotional state in the user that corresponds to *their* typical emotional state.

3 The SEMAINE Database

The SEMAINE database was recorded in order to provide training material for the speech and vision-based input components of the SEMAINE system. For this purpose, the functionality of the virtual agent system was imitated by a human operator using a Wizard-of-Oz scenario. Thus, users were encouraged to show emotions while naturally speaking about arbitrary topics.

The transcribed part of the database consists of 19 recordings with different English speaking users and has a total length of 6.2 h. Models used for the experiments in Section 5 are trained on recordings 1 to 10 (speech material from both, user and operator) and tested on recordings 11 to 19 (only speech from the user). The vocabulary size of the SEMAINE corpus is 3.4 k.

In addition to the SEMAINE database, two other spontaneous speech corpora were used for acoustic and language model training: the SAL corpus and the COSINE corpus. The SAL database was recorded under similar conditions as the SEMAINE corpus, which makes it well-suited for our application scenario. It has already been used in a large number of studies on emotional speech (for more details on the SAL database, see [10], for example). The COSINE corpus [11] contains multi-party conversations recorded in real world environments and is partly overlaid with indoor and outdoor noise sources. It consists of ten transcribed sessions with 11.4 h of speech from 37 different speakers and has a vocabulary size of 4.8 k.

4 Multi-stream LSTM-HMM

This section briefly outlines the multi-stream LSTM-HMM ASR system we use for enhanced keyword detection in emotionally colored speech (see Section 5.2). The main idea of this technique is to enable improved recognition accuracies by incorporating context-sensitive phoneme predictions generated by a Long Short-Term Memory network into the speech decoding process.

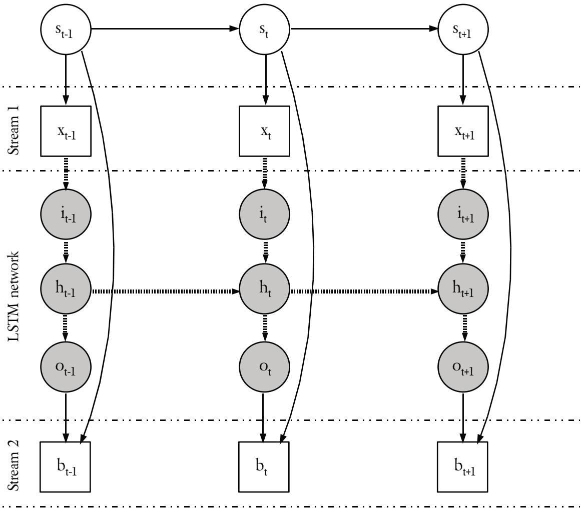


Fig. 1. Architecture of the multi-stream LSTM-HMM decoder: s_t: HMM state, x_t: acoustic feature vector, b_t: LSTM phoneme prediction feature, i_t, o_t, h_t: input, output, and hidden nodes of the LSTM network

LSTM networks [6] were introduced after the analysis of the error flow in conventional recurrent neural nets revealed that long range context is inaccessible to standard RNNs, since the backpropagated error either blows up or decays over time (vanishing gradient problem [5]). The LSTM principle is able to overcome the vanishing gradient problem and allows the network to learn the optimal amount of contextual information relevant for the classification task.

An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more memory cells, along with three multiplicative 'gate' units: the input, output, and forget gates. 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. The overall effect is to allow the network to store and retrieve information over long periods of time.

The structure of our multi-stream decoder can be seen in Figure 1: s_t and x_t represent the HMM state and the acoustic (MFCC) feature vector, respectively, while b_t corresponds to the discrete phoneme prediction of the LSTM network (shaded nodes). Squares denote observed nodes and white circles represent hidden nodes. In every time frame t the HMM uses two independent observations: the MFCC features x_t and the LSTM phoneme prediction feature b_t. The vector x_t also serves as input for the LSTM, whereas the size of the LSTM input layer i_t corresponds to the dimensionality of the acoustic feature vector. 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:

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

In every time step the LSTM generates a phoneme prediction according to Equation 1 and the HMM models $x_{1:T}$ and $b_{1:T}$ as two independent data streams. With $y_t = [x_t; b_t]$ being the joint feature vector consisting of continuous MFCC and discrete LSTM observations and the variable a denoting the stream weight of the first stream (i. e., the MFCC stream), the multi-stream HMM emission probability while being in a certain state s_t can be written as

$$p(y_t|s_t) = \left[\sum_{m=1}^{M} c_{s_t m}\mathcal{N}(x_t; \mu_{s_t m}, \Sigma_{s_t m})\right]^a \times p(b_t|s_t)^{2-a}. \tag{2}$$

Thus, the continuous MFCC observations are modeled via a mixture of M Gaussians per state while the LSTM prediction is modeled using a discrete probability distribution $p(b_t|s_t)$. The index m denotes the mixture component, $c_{s_t m}$ is the weight of the m'th Gaussian associated with state s_t, and $\mathcal{N}(\cdot; \mu, \Sigma)$ represents a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ. The distribution $p(b_t|s_t)$ is trained to model typical phoneme confusions that occur in the LSTM network. In our experiments, we restrict ourselves to the 15 most likely phoneme confusions per state and use a floor value of 0.01 for the remaining confusion likelihoods.

The applied real-time LSTM phoneme predictor is publicly available as part of our on-line speech feature extraction engine openSMILE [12].

5 Experiments and Results

In the following we will show the effects of using multi-condition training for a keyword detector based on a conventional single-stream continuous ASR system (Section 5.1), and the performance gain that can be obtained when applying the multi-stream LSTM-HMM principle (Section 5.2).

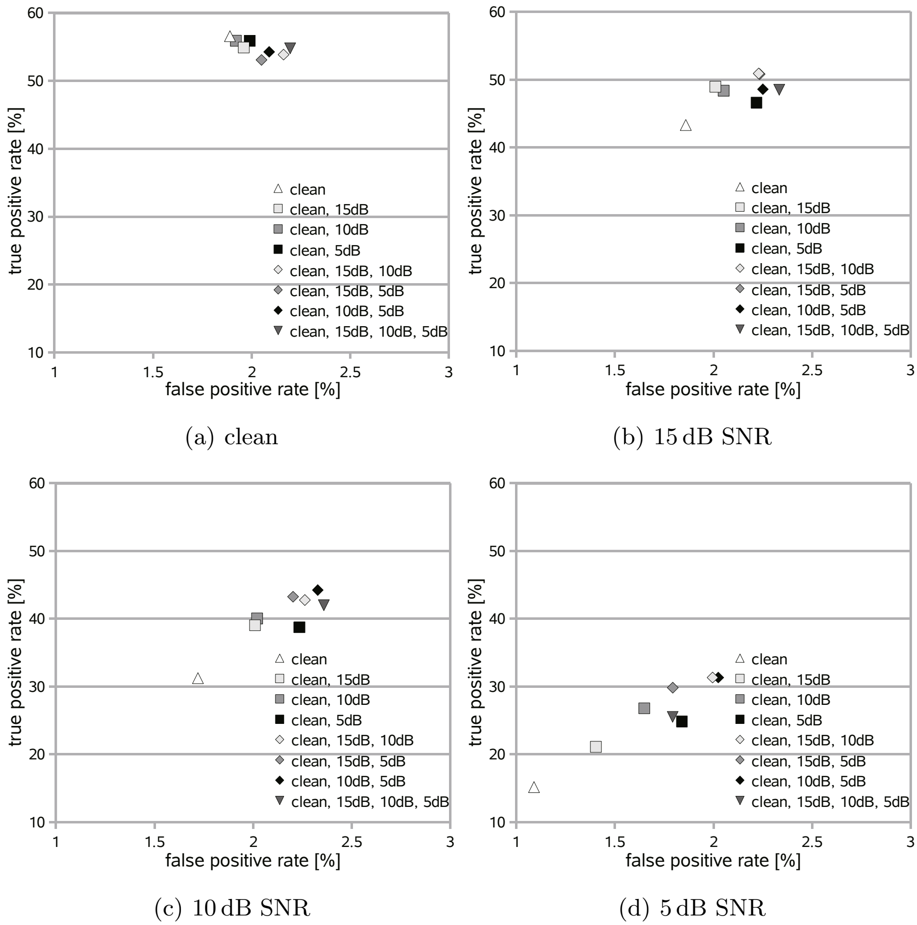
5.1 Multi-condition Training

To improve keyword detection accuracy in noisy conditions, we investigated true positive and false positive rates when including noisy speech material in the training process. For all experiments, a part of the training material consisted of unprocessed versions of the SEMAINE database (recordings 1 to 10), the SAL corpus, and the COSINE database. This speech material will be referred to as *clean* in the ongoing (even though the COSINE corpus was partly recorded under noisy conditions). In addition to the 'clean' models, we evaluated different extensions of the training material by adding distorted versions of the SEMAINE and the SAL corpus. For this purpose, we superposed the clean speech with additive noise at different SNR levels: 15 dB, 10 dB, and 5 dB. We considered both, white Gaussian noise and babble noise from the NOISEX database. For evaluation, we used clean and distorted versions of the SEMAINE database (recordings 11 to 19). Since conversational agents such as the SEMAINE system are often used while other people talk in the background, the babble noise evaluation scenario is most relevant for our application. We considered a set of 173 keywords and three different non-linguistic vocalizations (*breathing*, *laughing*, and *sighing*). The training/test distribution for *breathing*, *laughing*, and *sighing* was 124/54, 268/227, and 45/8, respectively. Keyword detection was based on simply searching for the respective words in the most likely ASR hypothesis. The applied trigram language model was trained on the SEMAINE corpus (recordings 1 to 10), the SAL database, and the COSINE database (total vocabulary size 6.1 k). Via openSMILE [12], 13 cepstral mean normalized MFCC features along with first and second order temporal derivatives were extracted from the speech signals every 10 ms. All cross-word triphone HMMs consisted of 3 emitting states with 16 Gaussian mixtures per state. For non-linguistic vocalizations, we trained HMMs consisting of 9 states.

Figures 2(a) to 2(d) show the Receiver Operating Characteristic (ROC) operating points for clean test material as well as for speech superposed with babble noise at 15 dB, 10 dB, and 5 dB SNR, respectively, when using different acoustic models. As can be seen in Figure 2(a), models exclusively trained on clean speech lead to the best performance for clean test data. We obtain a true positive rate of 56.58 % at a false positive rate of 1.89 % which is in the range of typical recognition rates for highly disfluent, spontaneous, and emotionally colored speech [7]. Including noisy training material slightly increases the false positive rate to up to 2.20 % at a small decrease of true positive rates. Yet, when evaluating the models on speech superposed by babble noise, multi-condition training significantly increases the true positive rates. A good compromise between high true positive rates and low false positive rates in noisy conditions can be obtained by applying the acoustic models denoted as 'clean, 15 dB, 10 dB' in Figures 2(a) to 2(d), i. e., models trained on the clean versions of the SEMAINE, SAL, and COSINE corpus, on the SEMAINE and SAL database superposed by babble noise at 15 dB SNR, and on the 10 dB versions of the SEMAINE and SAL database. For test data superposed by babble noise, this training set combination leads to the highest average true positive rate (41.66 %, see Table 1) at a tolerable average false positive rate. A similar result can be observed for the evaluation on test data corrupted by white noise (see Table 2). Models that are partly trained on speech superposed by white noise enable higher true positive rates in noisy conditions than 'clean' models. As for the babble noise scenario, a combination of clean, 15 dB SNR, and 10 dB SNR training data results in the best true positive/false positive compromise.

5.2 Multi-stream Decoding

To improve keyword detection in clean conditions, we implemented and evaluated the multi-stream LSTM-HMM decoder introduced in Section 4. Since the LSTM network was trained on framewise phoneme targets, we used an HMM system to obtain phoneme borders via forced alignment. The multi-stream system was trained on the clean versions of the SEMAINE, SAL, and COSINE databases



(a) clean

(b) 15 dB SNR

(c) 10 dB SNR

(d) 5 dB SNR

Fig. 2. ROC operating points obtained for different acoustic models when tested on clean speech and speech superposed by babble noise at 15, 10, and 5 dB SNR; acoustic models were trained on unprocessed versions of the SEMAINE, SAL, and COSINE corpus (‘clean’) and on noisy versions of the SEMAINE and SAL corpus using different SNR level combinations (babble noise)

and applied an LSTM network with a hidden layer consisting of 128 memory blocks. Each memory block contained one memory cell.

For LSTM network training we used a learning rate of 10^{-5} and a momentum of 0.9. Prior to training, all weights were randomly initialized in the range from -0.1 to 0.1. Input and output gates used tanh activation functions, while the forget gates had logistic activation functions. We trained the networks on the standard (CMU) set of 41 different English phonemes, including targets for *silence*, *breathing*, *laughing*, and *sighing*. The stream weight variable a was set to one.

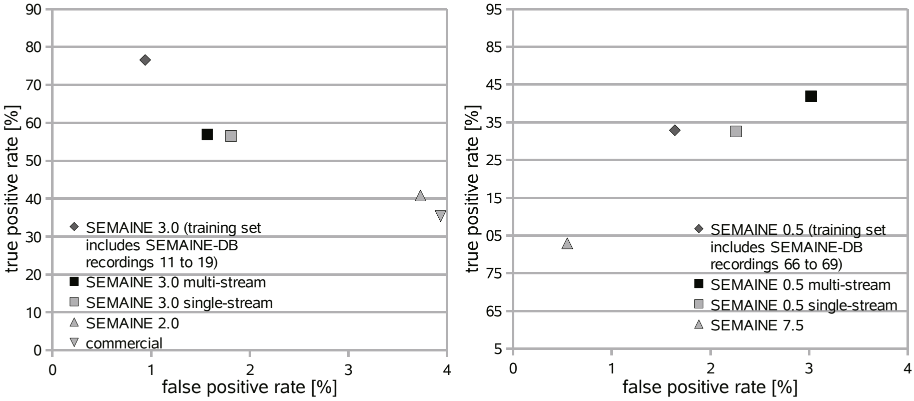
Table 1. Babble noise: Average true positive rates (tpr) and false positive rates (fpr) obtained with acoustic models trained on clean data and speech superposed by babble noise at different SNR conditions; clean and noisy test condition

training data	**test condition**			
SNR used for superposition	**babble noise**		**clean**	
with babble noise	**tpr [%]**	**fpr [%]**	**tpr [%]**	**fpr [%]**
clean	29.89	1.56	56.58	1.89
clean, 15 dB	36.37	1.81	54.91	1.96
clean, 10 dB	38.40	1.91	55.92	1.92
clean, 5 dB	36.73	2.10	55.90	1.99
clean, 15 dB, 10 dB	**41.66**	2.16	53.87	2.16
clean, 15 dB, 5 dB	41.29	2.08	53.08	2.05
clean, 10 dB, 5 dB	41.38	2.20	54.28	2.09
clean, 15 dB, 10 dB, 5 dB	38.67	2.16	54.79	2.20

Table 2. White noise: Average true positive rates (tpr) and false positive rates (fpr) obtained with acoustic models trained on clean data and speech superposed by white noise at different SNR conditions; clean and noisy test condition

training data	**test condition**			
SNR used for superposition	**white noise**		**clean**	
with white noise	**tpr [%]**	**fpr [%]**	**tpr [%]**	**fpr [%]**
clean	19.81	1.26	56.58	1.89
clean, 15 dB	39.40	2.28	57.31	2.06
clean, 10 dB	39.33	2.44	56.50	2.03
clean, 5 dB	23.65	1.20	56.02	2.01
clean, 15 dB, 10 dB	**42.47**	2.54	54.55	2.21
clean, 15 dB, 5 dB	42.11	2.57	54.28	2.16
clean, 10 dB, 5 dB	41.27	2.60	54.09	2.04
clean, 15 dB, 10 dB, 5 dB	**42.48**	2.69	50.42	2.27

The ROC operating points representing the keyword detection performance of the standard HMM (SEMAINE 3.0 single-stream) and the LSTM-HMM (SEMAINE 3.0 multi-stream) can be seen in Figure 3(a). All systems were evaluated on recordings 11 to 19 from the SEMAINE database. At a slight increase of the true positive rate, the incorporation of LSTM phoneme predictions can significantly reduce the false positive rate from 1.89 % to 1.57 %. For comparison, we also included the results for a preliminary version of the SEMAINE keyword detector (referred to as the SEMAINE 2.0 system [9]) which does not apply an in-domain language model and thus cannot compete with the current version (SEMAINE 3.0). Figure 3(a) also shows the performance obtained with a commercial recognizer as used in [13]. The comparably low performance of the commercial system indicates that using acoustic models tailored for the recognition of emotionally colored speech is essential for virtual agent applications such as the SEMAINE system. Since the final SEMAINE 3.0 keyword detector is trained on the *whole* SEMAINE database (including recordings 11 to 19),



(a) keyword detection (b) non-linguistic vocalization detection

Fig. 3. ROC operating points obtained for different variants of the SEMAINE keyword and non-linguistic vocalization detector

Figure 3(a) also shows the ROC performance obtained with models trained on all SEMAINE data. Note, however, that this configuration does not allow for a realistic performance assessment since training and test sets are not disjoint in this case. The reliability of non-linguistic vocalization detection (i. e., recognizing the events *breathing*, *laughing*, and *sighing*) can be seen in Figure 3(b): Again the multi-stream approach leads to a higher true positive rate, however, – in contrast to the keyword detection experiment – at the expense of a higher false positive rate.

6 Conclusion

This paper investigated how a keyword detector incorporated in a conversational agent system can be improved via multi-stream LSTM-HMM decoding and multi-condition training. We proposed a multi-stream system that models context-sensitive phoneme predictions generated by a Long Short-Term Memory network. In conformance with our previous observations concerning LSTM-based keyword spotting [7], we found that the LSTM principle is well-suited for robust phoneme prediction in challenging ASR scenarios. Performance gains in noisy conditions could be obtained applying multi-condition training. Since virtual agents are often used while people talk in the background, we mainly considered test conditions during which the speech signal is superposed by babble noise. Incorporating training material that is overlaid by background voices at different SNR conditions could enhance the noise robustness of keyword detection.

To further improve multi-stream LSTM-HMM keyword detection for conversational agents, future experiments should evaluate alternative network topologies such as *bottleneck* LSTM architectures as well as bidirectional context modeling for refinement of sentence hypotheses at the end of an utterance.

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

References

1. McTear, M.F.: Spoken dialogue technology: enabling the conversational user interface. ACM Computing Surverys 34(1), 90–169 (2002)
2. Droppo, J., Acero, A.: Environmental robustness. In: Handbook of Speech Processing, pp. 658–659. Springer, Heidelberg (2007)
3. Schuller, B., Wöllmer, M., Moosmayr, T., Rigoll, G.: Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement. Journal on Audio, Speech, and Music Processing (2009), ID 942617
4. Zhu, Q., Chen, B., Morgan, N., Stolcke, A.: Tandem connectionist feature extraction for conversational speech recognition. In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 223–231. Springer, Heidelberg (2005)
5. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Kremer, S.C., Kolen, J.F. (eds.) A Field Guide to Dynamical Recurrent Neural Networks, pp. 1–15. IEEE Press, Los Alamitos (2001)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997)
7. Wöllmer, M., Eyben, F., Graves, A., Schuller, B., Rigoll, G.: Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. Cognitive Computation 2(3), 180–190 (2010)
8. Wöllmer, M., Eyben, F., Schuller, B., Rigoll, G.: Recognition of spontaneous conversational speech using long short-term memory phoneme predictions. In: Proc. of Interspeech, Makuhari, Japan, pp. 1946–1949 (2010)
9. Schröder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C., Schuller, B.: Towards responsive sensitive artificial listeners. In: Proc. of 4th Intern. Workshop on Human-Computer Conversation, Bellagio, Italy, pp. 1–6 (2008)
10. Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G.: Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE Journal of Selected Topics in Signal Processing 4(5), 867–881 (2010)
11. Stupakov, A., Hanusa, E., Bilmes, J., Fox, D.: COSINE - a corpus of multi-party conversational speech in noisy environments. In: Proc. of ICASSP, Taipei, Taiwan (2009)
12. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE - the Munich versatile and fast open-source audio feature extractor. In: Proc. of ACM Multimedia, Firenze, Italy, pp. 1459–1462 (2010)
13. Principi, E., Cifani, S., Rocchi, C., Squartini, S., Piazza, F.: Keyword spotting based system for conversation fostering in tabletop scenarios: preliminary evaluation. In: Proc. of HSI, Catania, Italy, pp. 216–219 (2009)