Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

How a Listener Influences the Speaker

Timo Baumann

Universität Hamburg
Department of Informatics
baumann@informatik.uni-hamburg.de

Abstract

Listeners typically provide feedback while listening to a speaker in conversation and thereby engage in the *co-construction* of the interaction. We analyze the influence of the listener on the speaker by investigating how her verbal feedback signals help in modeling the speaker's language. We find that feedback from the listener may help in modeling the speaker's language, whether through the listener's feedback as transcribed, or the acoustic signal directly. We find the largest positive effects for end of sentence as well as for pauses mid-utterance, but also effects that indicate we successfully model elaborations of ongoing utterances that may result from the presence or absence of listener feedback.

Index Terms: conversation, feedback, language model, RNNLM, CNN

1. INTRODUCTION

Conversations have long been understood as being co-constructed as a shared activity between speaker and listener [1]. While most obviously, successful dialog consists of contributions (*turns*) that are exchanged between interlocutors and that relate to and build on each other in the form of adjacency pairs [2], it appears that influences can also be more fine-grained. Rather than merely on a turn-by-turn basis in which one participant's utterance influences the other participant's next utterance, feedback can be exchanged more quickly and integrated immediately into the conversation, as human dialog participants perceive and understand continuously [3] and are able to adapt ongoing speech with low latency [4].

Yngve [5] coined the term "back-channel" utterances for listener behaviours that overlap with ongoing speech on the 'primary channel' (without being interpreted as out-of-place or as interrupting an ongoing turn [6]). Such utterances can be short feedback words like 'yeah', 'good', 'right', conversational grunts [7], audible inhalation, yawning, lengthening or softening of delivery in the primary channel, and so on. In face-to-face communication, feedback signals extend beyond the audible modality and include gaze, facial expressions, gestures, and posture shifts [8].

Speakers integrate feedback on the back-channel (or the absence of feedback where it would be expected) into their ongoing turn on the forward channel. Thus, the presence or absence of feedback where it would be appropriate, and its realization may be informative in modeling the next word selection decisions of the speaker *during* the ongoing utterance.

Spoken dialog systems obviously make use of the words in the most recent user utterance (and their meaning in terms of a dialog act) *after* the utterance is completed, to determine the next system actions. Recently, RNN-based language models [9] have been improved by integrating the previous speaker's words and/or an externally provided dialog act tag when determining the words of the upcoming speaker [10], yielding a 3 % improvement of perplexity over the baseline. This technique has also been shown to improve speech recognition rescoring by 1-4 % relative word error reduction [11]. In contrast to that work, which uses high-level context that precedes the turn in question, we here intend to use low-level context that is acquired during the turn, which likely is orthogonal. The speaker's acoustic/prosodic information has also previously been shown to improve a language model [12]. This work, to our knowledge, is the first to consider such information coming from the *listener*.

In human-computer interaction, a recent system that adapts its utterances to user feedback [13] was rated as more helpful and understanding in the conversation than a baseline system. The system uses a human operator to encode the user-provided feedback and the system's generation is performed by an adaptive extension [14] of a rule-based generation component [15] limited to a small domain. We see our research as a bridge from such rule-based systems with few, explicit, and expert-defined feedback responses for adaptive behaviour in small domains towards trainable open-domain NLG approaches based on encoder-decoder neural networks [16] that take listener feedback into account. As a step towards this goal, we analyze the influence of listening behaviour on speaker modeling with RNNs in this paper.

In the following, we detail our idea and the neural architecture that we use in Sections 2 and 3, before we describe the corpus that we use and the experiments that we perform in Section 4. We also performed a thorough manual analysis of perplexities word-by-word and report some interesting findings in Section 5 and conclude in Section 6.

2. BASIC IDEA

The assumption underlying our model is that speakers are able to consider the listeners' behaviours and to adapt their own speech accordingly with little delay. An example can be found in Fig. 1: the speaker (A) delivers a potentially controversial assertion ("that's been the big role of government") to which he amends "I guess" to clearly state that this assertion is up for discussion. He then audibly breathes in, which, however, does not trigger the turn to be taken or feedback to be produced by B. He then starts an extension of his assertion ("I mean") which he promptly aborts when B reacts ("well, uh, it's supposed to . . . ") and instead amends "generally" which can be seen as further softening his argument. Also, note how B interupts himself during this amendment and waits for A to finish.

In this work, we are particularly interested in this within-utterance interplay of the interlocutors (who are both speakers and listeners). For each speaker, we will investigate the influence

This work was supported in part by a Google cloud research award. The author would like to thank Tom Mitchell for valuable discussions and support, as well as three anonymous reviewers for helpful comments.

10.21437/SpeechProsody.2020-198

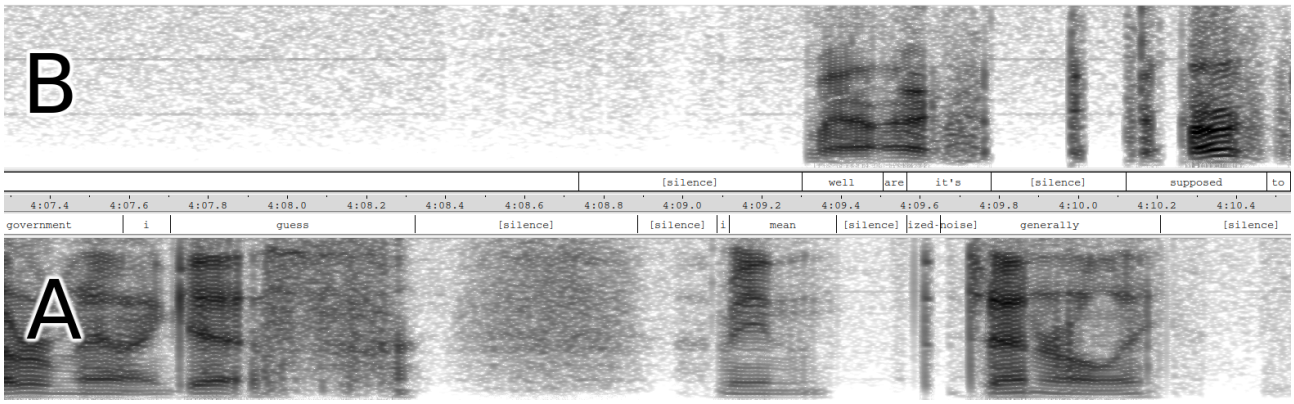

Figure 1: *Example of an utterance extension: "[that's been the big role of] government I guess ... I mean ... ‹noise› generally ‹/S›" (sw4572) shaped by (a) lack of feedback (triggering "I mean"), and (b) negative feedback (triggering "generally").*

of the other's behaviour in terms of: (a) the words or conversational grunts spoken *as transcribed* that were completed before the word to be predicted starts, and (b) the *direct parameterization* of the listener acoustics preceding the word to be predicted.

We expect a direct model based on the listener acoustics to be more powerful as it will be able to condition on more subtle details like breath, take into account the prosody and thus differentiate feedback signals. It will also be easier to apply given that it does not need to rely on just-in-time speech recognition results for the listener but can be fed with low delay from raw audio. The complex speech signal, however, is difficult to model and a suitable representation may need a lot of training material. Thus, our direct, acoustics-based model may be impeded by issues related to data sparsity.

3. ARCHITECTURE

The complete architecture for our model is depicted in Fig. 2. A neural sequence model encodes what has been spoken so far by the speaker, based on an initial state that possibly takes into account some prior context (such as the previous user utterance for a dialog LM [10], and/or a representation of what is to be spoken for NLG [17]). The estimation layer generates full probability distributions over the vocabulary from the sequence model's internal state. Words are input into the model as embeddings that are estimated during training. We set the embedding size to 128 dimensions.[1]

Listener feedback is integrated into the model's estimation layer as it becomes available during decoding. We have chosen a *late fusion* approach in which integration occurs in the estimation layer, rather than via the recurrent model.

For our **text**-based models, we use the listener feedback as transcribed in the corpus, i.e., the word most recently completed by the listener, via the same embeddings as the speaker's words. Most of the time, the listener feedback is '[silence]'.

For our **audio**-based models, we use a two-layer convolutional neural network on the most recent listener audio as represented by a spectrogram. Our represention is limited by memory and trainability. An additional concern was a uniform size of convolution and pooling filters on both layers. After initial testing, we settled on using the most recent 68 frames (680 ms) of listener acoustics (the largest size we could fit into memory), passed through a log-melfilterbank that yields 26 spectral dimensions [18] which are then z-normalized. We feed these features into a two-layer convolutional neural network which uses $3_f \times 5_t$ convolutional filters (4 on first, 16 on second layer), $2_f \times 4_t$ max-pooling and ELU non-linearities [19], yielding a 240-dimensional representation.

We also use a combination of text and audio as **text+audio**.

[1]Note that changes in the baseline model, for example pre-estimated embeddings, would likely carry over to the listener-enhanced models.

4. EXPERIMENT

We use the Switchboard corpus [20] with the ISIP transcriptions [21] and time-alignments. Time-alignments are used to determine the most recently spoken listener word and/or the listener's most recent audio. We adopt the same training/validation/test splits as [22] which ensure that each speaker appears only in one of the sets. We re-combine utterances that are only interrupted by a short intervening back-channel of maximally 1.3 s into one. We limit the vocabulary to those 14k tokens that occur at least 2 times and replace all others with ‹unk›.

In this work, we are primarily interested in finding out whether listener feedback can be fruitfully used to model the speaker's language. We therefore compare a number of listener-enhanced models with their respective baselines. We start with a comparison of unigram models (i.e., completely ignoring the speaker's word history), via bigrams (only the most recent speaker's history), to the full RNN model depicted in Fig. 2 (using LSTM cells [23]). We take this approach in order to tease apart the merit from using listener information vs. the possiblity that the large number of parameters for the full model cannot successfully be trained on the limited corpus.

Our models are implemented in dyNet [24] and our baseline

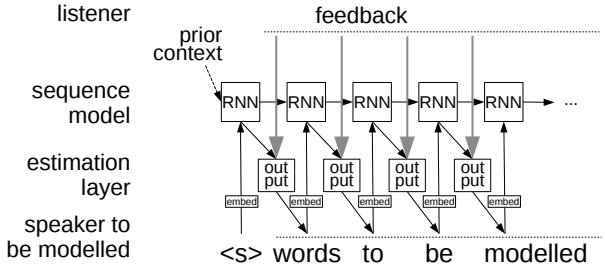

Figure 2: *Architecture of our model that incorporates the listener.*

unigram/bigram models yield very similar results to SRILM [25], indicating that these models can be optimally trained on the dataset. We train for a maximum of 15 iterations using AdaGrad [26] using a mini-batch size of 32, and keep the best models as determined on the validation set, for which we also report the test performance. We use dropout of $p = 0.2$ on the fully-connected and RNN layers. The RNN states are 128 dimensional. None of these parameters have been heavily optimized but should be seen as moderately informed best guesses.

The code that can be used to replicate our results can be found at `http://bitbucket.org/timobaumann/spepro2020`.

5. RESULTS & ANALYSIS

The language modeling performance of our models is reported in Table 1 in terms of cross-entropy.[2] Conditioning the unigram model on the listener's most recently completed word leads to an improvement of about 0.07 bit (a 5 % reduction in perplexity), which is already highly significant (per-word sign-test, $p < .001$). Conditioning the unigram model on the most recent audio in the listener's channel leads to a (larger) improvement and with both the listener's audio and transcribed words, the improvement amounts to 0.23 bit, an overall 15 % reduction in perplexity.

Improvements for bigram-based models are much smaller. We speculate that (a) the previous word contains similar contextual information as the listener feedback (e.g., sentence start markers and overlapping turns, or feedback-inviting words and the actual feedback). Furthermore, (b) some Switchboard recordings contain a significant amount of crosstalk between audio channels. As a result, the unigram model's acoustic part may actually focus on the crosstalked speaker's speech (making this approach effectively a bigram model). However, this effect can be excluded for the gains of both the text-based conditions as well as the bigram conditions.

We hence estimate the contribution of the listener in modeling the speaker to around 0.1 – 0.2 bit (the equivalent of a ~10 % reduction in perplexity). Overall, it appears that listener audio is more informative than transcribed text and that their combination is partially cumulative.

We have also attempted to integrate listener feedback into the RNN-based LM, with moderate results. While we do see some improvement of 0.02–0.04 bit for the text+audio condition, results deteriorate for the text and audio-only conditions. We believe that our model at present is unable to leverage the full potential of listener feedback, as it is quite simple in many ways: None of the hyper-parameters of the models have been tuned which could potentially yield some gains. Additionally, joint embeddings for the speaker's and the listener's words may have been a bad choice and could explain the lacking performance of the models that include listener text. Also, we observed very quick convergence (and overfitting) for the CNN-based part of the model vs. much slower learning for the RNN-based part. With both parts of the model sharing the same learning rate, by the time the RNN is sufficiently well trained, the CNN does not generalize well anymore. This could be mitigated by more advanced training procedures and/or more training data.

[2]Cross-entropy is the dual logarithm of perplexity and yields the information in bit. Results can be compared via subtraction of cross-entropy, whereas perplexity must be compared as ratios, which is less straightforward.

Table 1: *Modeling performance on Switchboard in terms of cross-entropy in bits (lower is better), as well as gain over baseline (for listener-enhanced models; higher is better).*

	validation		test	
model		gain		gain
unigram	8.15		8.24	
+ listener text	8.06	0.09	8.17	0.07
+ listener audio	7.94	0.21	8.07	0.17
+ listener text+audio	7.87	0.38	8.01	0.23
bigram	6.52		6.66	
+ listener text	6.49	0.03	6.64	0.02
+ listener audio	6.50	0.02	6.65	0.01
+ listener text+audio	6.49	0.03	6.64	0.02
RNN	6.13		6.28	
+ listener text	6.15	-0.02	6.35	-0.07
+ listener audio	6.20	-0.03	6.36	-0.08
+ listener text+audio	6.09	0.04	6.26	0.02

5.1. Detailed Analysis

We perform an analysis of the per-word performance of the listener-enhanced vs. baseline model in terms of how well each predicts the next word in the utterance to find out *how* and *where* the listener influences the speaker (and to justify the paper title). For the numbers reported below, we focus on the text+audio model in the bigram condition.

We find (a) markers that denote silence ('[silence]', '‹/S›') are those high-frequency tokens that are most considerably boosted (with a median improvement above 0.5 bit). In other words: our model is very successful in capturing when the speaker will stop or temporarily interrupt the turn. Boosts are particularly high for interruptions, as shown by the bigrams 'to [silence]', 'a [silence]', or 'the [silence]' with well above 1 bit improvement. Likewise, (b) '‹S›' and '[silence]' are among the frequent preceding speaker tokens that yield large improvements, indicating good turn-initial performance and improvements after pauses. (c) Backchannel utterances in particular are boosted, as indicated by the bigrams '‹S› um-hum', '‹S› uh-huh'.

(d) In terms of phenomena likely to appear mid-utterance (i.e., what we are particularly interested about), we do find improvements for markers of low-latency grounding used by the speaker (e.g., 'you know'), and feedback to a brief interruption ('[silence] yeah'). Regarding material that is spoken (by the listener) immediately before the speaker's word to be modelled, we find that back-channels ('uh-huh', 'um-hum', 'yeah', 'right') as well as words that can be used to hand over or deny a turn ('so', 'well') lead to significant and large (~0.3 bit) improvements.

(e) It is more difficult to aggregate statistics about high-level situations in which the proposed model performs better. To this end, we manually checked numerous turns and their contexts with the largest average improvements (excluding very short turns which most often contain back-channels). We find that large improvements often occur during the start or end of the interaction, often for both speakers, which may be due to the tight and highly conventionalized temporal co-alignment of both interlocutors during this phase [27]. Indicative of the model doing what it's supposed to, we also find gains for most of the words in Fig. 1, as well as for many other cases of utterance interruptions and extensions.

As a final note: while all the observations reported in this

analysis are backed by significant differences between the two models, we need to stress that random differences will occur between different training runs of the same model architecture (in the individual *per-word* performance – they equal out in the overall model performance). Thus, although only those results have been reported that also intuitively seem plausible, the abovementioned results must be taken with a grain of salt.

6. SUMMARY AND FUTURE WORK

This paper explores whether a model of the speaker's words profits from including information about the listener's verbal behaviour *while* the speaker is speaking. We have found that indeed, this information helps in modeling the speaker, with a perplexity improvement of 15 % for unigrams but only a few percent for the RNN-based LM. Some issues of integrating listener feedback with the RNN-based LM may be related to ineffective training with the unoptimized acoustic model overfitting more quickly than the RNN-based LM. Beyond better training methods, more advanced acoustic feature representations that yield much better normalization than what we use and are pre-trained on additional data [28] may help to overcome this issue.

We believe our findings could help in dialog applications, for example speech recognition (as in [11]) for call-center agent support [29]. Another interesting extension would be towards multi-party conversations for example in meeting transcription [30]. Meetings feature a large proportion of overlaps and feedback which our method exploits; at the same time, one would have to model multiple listeners which poses interesting problems.

Overall, our models have a wide range of hyper-parameters, in particular the parameters for the listening acoustics, and we have not performed any tuning but merely used a combination of best guesses made fit with memory and computation time constraints. It would be interesting to investigate what range of acoustic context is most useful (longer/shorter) and whether leaving out the most recent context (say, 100 ms) impacts performance given that very late information can hardly be integrated in time by the speaker.

We have tried to exclude the influence of the *speaker*'s acoustics in modeling his words, as this would most likely already be dealt with by the acoustic model in speech recognition. However, depending on the ASR architecture, integrating the acoustic history into the language model (which overall, is responsible for managing history in ASR) might also be a direction worth exploring; see also [12].

The main reason for excluding the speaker, though, is that we want to extend our approach towards responsive NLG for an attentive virtual agent. In fact, our language models can be directly extended to the encoder-decoder framework by using an initial state that derives from an encoding of the logical form of the utterance to be generated. Decoding of the NLG could then be performed time-synchronously to speech production (and using incremental speech synthesis [31]) and consider the most recent listener behaviours.

Finally, a lot of feedback behaviours such as nods or posture are not communicated via the audio channel. While we have based our research on telephone conversations in which the listener knows that only verbal feedback can be observed by the speaker, we have only modelled a small subset of all possible listener feedbacks. It would be extremely interesting to include video analysis of listeners into an automated broad-coverage model for speaker modeling in situated dialog in the future.

7. References

[1] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[2] E. A. Schegloff, "Sequencing in conversational openings," *American Anthropologist*, vol. 70, no. 6, 1968.

[3] M. K. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy, "Integration of visual and linguistic information in spoken language comprehension," *Science*, vol. 268, no. 5217, pp. 1632–1634, 1995.

[4] W. J. Levelt, *Speaking: From Intention to Articulation*. MIT Press, 1989.

[5] V. H. Yngve, "On getting a word in edgewise," in *Chicago Linguistics Society, 6th Meeting*, 1970, pp. 567–578.

[6] S. Duncan, Jr and G. Niederehe, "On signalling that it's your turn to speak," *Journal of Experimental Social Psychology*, vol. 10, no. 3, pp. 234–247, 1974.

[7] N. Ward, "Non-lexical conversational sounds in American English," *Pragmatics & Cognition*, vol. 14, no. 1, pp. 129–182, 2006.

[8] J. M. Wiemann and M. L. Knapp, "Turn-taking in conversations," *Journal of Communication*, vol. 25, no. 2, pp. 75–92, 1975.

[9] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of INTERSPEECH*, no. 9, 2010, pp. 1045–1048.

[10] B. Liu and I. Lane, "Dialog context language modeling with recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE Int. Conf. on*, 2017, pp. 5715–5719.

[11] W. Xiong, L. Wu, J. Zhang, and A. Stolcke, "Session-level language modeling for conversational speech," in *Proceedings EMNLP*, 2018.

[12] N. G. Ward, A. Vega, and T. Baumann, "Prosodic and temporal features for language modeling for dialog," *Speech Communication*, vol. 54, no. 2, pp. 161–174, 2012.

[13] H. Buschmeier and S. Kopp, "Communicative listener feedback in human–agent interaction: artificial speakers need to be attentive and adaptive," in *Proceedings of the 17th Int. Conf. on Autonomous Agents and Multiagent Systems*, 2018, pp. 1213–1221.

[14] H. Buschmeier, T. Baumann, B. Dorsch, S. Kopp, and D. Schlangen, "Combining incremental language generation and incremental speech synthesis for adaptive information presentation," in *Proceedings of SigDial*, 2012, pp. 295–303.

[15] M. Stone, C. Doran, B. Webber, T. Bleam, and M. Palmer, "Microplanning with communicative intentions: The SPUD system," *Computational Intelligence*, vol. 19, pp. 311–381, 2003.

[16] O. Dušek, J. Novikova, and V. Rieser, "Findings of the E2E NLG challenge," in *Proceedings of the 11th International Conference on Natural Language Generation*, 2018, pp. 322–328.

[17] O. Dušek and F. Jurcicek, "A context-aware natural language generator for dialogue systems," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, pp. 185–190.

[18] J. Lyons, D. Y.-B. Wang, Gianluca, H. Shteingart, E. Mavrinac, Y. Gaurkar, W. Watcharawisetkul, S. Birch, L. Zhihe, J. Hölzl, J. Lesinskis, H. Almér, C. Lord, and A. Stark, "jameslyons/python_speech_features: release v0.6.1," Jan. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3607820

[19] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *CoRR*, vol. abs/1511.07289, 2015.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE Int. Conf. on*, vol. 1, 1992, pp. 517–520.

[21] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of Switchboard," in *Fifth International Conference on Spoken Language Processing*, 1998.

[22] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] G. Neubig, C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn, K. Duh, M. Faruqui, C. Gan, D. Garrette, Y. Ji, L. Kong, A. Kuncoro, G. Kumar, C. Malaviya, P. Michel, Y. Oda, M. Richardson, N. Saphra, S. Swayamdipta, and P. Yin, "Dynet: The dynamic neural network toolkit," *arXiv preprint arXiv:1701.03980*, 2017.

[25] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at sixteen: Update and outlook," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, p. 5.

[26] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[27] G. Jefferson, "A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences," *Semiotica*, vol. 9, no. 1, pp. 47–96, 1973.

[28] B. Milde and C. Biemann, "Unspeech: Unsupervised speech context embeddings," in *Proceedings of Interspeech*, 2018, pp. 2693–2697.

[29] A. T. Farrell, "Call centre agent automated assistance," Patent US 6,721,416 B1, 2004-04-13.

[30] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, "The AMI meeting transcription system: Progress and performance," in *International Workshop on Machine Learning for Multimodal Interaction*, 2006, pp. 419–431.

[31] T. Baumann and D. Schlangen, "INPRO_iSS: A component for just-in-time incremental speech synthesis," in *Proceedings of ACL System Demonstrations*, 2012.