Language Modeling of Nonverbal Vocalizations in Spontaneous Speech

Dmytro Prylipko[1], Bogdan Vlasenko[1], Andreas Stolcke[2], and Andreas Wendemuth[1]

[1] Cognitive Systems, Otto-von-Guericke University, 39016 Magdeburg, Germany
{dmytro.prylipko,bogdan.vlasenko,andreas.wendemuth}@ovgu.de
[2] Conversational Systems Lab, Microsoft, Mountain View, CA, USA
andreas.stolcke@microsoft.com

Abstract. Nonverbal vocalizations are one of the characteristics of spontaneous speech distinguishing it from written text. These phenomena are sometimes regarded as a problem in language and acoustic modeling. However, vocalizations such as filled pauses enhance language models at the local level and serve some additional functions (marking linguistic boundaries, signaling hesitation). In this paper we investigate a wider range of nonverbals and investigate their potential for language modeling of conversational speech, and compare different modeling approaches. We find that all nonverbal sounds, with the exception of breath, have little effect on the overall results. Due to its specific nature, as well as its frequency in the data, modeling of breath as a regular language model event leads to a substantial improvement in both perplexity and speech recognition accuracy.

1 Introduction

The conversational speech we produce every day is remarkably different from the types of speech for which language technology is successfully applied [1]. Campbell argues that spontaneous speech encodes two distinct streams of information, linguistic and interpersonal [2]. While the former carries the verbal meaning of a message, the latter accompanies it and enriches our speech with paralinguistic messages and cues. They coexist and merge into a common information stream which is called human spoken speech. For recognition and synthesis of spontaneous speech, special attention must be paid to conversational speech phenomena, both in acoustic and language modeling.

An example of such phenomena are nonverbal and nonspeech sounds. Nonverbals produced with human vocal apparatus can provide valuable paralinguistic cues, but do not have any linguistic meaning (filled pauses, clicks and lip smacks, cough, laughter). They occur frequently in spontaneous speech and have been investigated since 1990s.

Schultz and Rogina in [3] took into consideration different human and non-human noises present in transcriptions and found that modeling them as regular words improves recognition performance compared with treating them as silences, i.e. non-linguistic events expose some regularities well-captured by standard N-grams.

Further research has been devoted mostly to the use of filled pauses and linguistic *disfluencies* (DFs) such as false starts, repetitions, or repairs. Stolcke and Shriberg introduced a language model (LM) that explicitly models these phenomena [4]. That work confirmed that DFs have a systematic and nonrandom nature that can be captured

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 488–495, 2012.

Table 1. Distribution and type of the nonverbals in the Verbmobil corpus

Filled pauses	Breath	Human noise	Laugh	Throat clean	Swallow	Lip smack	**Total**
8 464	21 149	1 710	231	93	228	5 844	**37 719**

by LMs and contain information for predicting following words. They also argue that one function of filled pauses is marking the beginning of the linguistic segment. Further research confirmed importance of keeping filled pauses in the LM conditioning context and their role as linguistic boundaries [5], [6]. These articles address two of the four fundamental challenges outlined by Shriberg in [7], namely, recovering hidden punctuation (such as sentence and phrase boundaries) and coping with disfluencies.

The results with filled pauses show us the potential of using spontaneous speech characteristics (usually regarded as a source of problems) for modeling more natural human speech. However, among all nonverbal sounds, only the role of filled pauses has been investigated extensively for LM purposes.

Modern transcriptions of spontaneous speech corpora contain a broad range of different nonverbal and nonspeech markers such as breath, laughter, cough, lip smack, verbal and nonverbal noises, throat cleaning etc. But it is not obvious whether it is worth modeling them as LM events or would be better to eliminate them from the model. Furthermore, state-of-the-art language modeling tools provide a wide range of modeling techniques, some specially designed for speech disfluencies. In this paper we investigate these techniques applied to several classes of nonverbal sounds, in order to determine which nonverbals have the potential to improve language models of conversational speech.

2 Speech Corpus Overview

We performed our experiments on the Verbmobil corpus (German part) [8]. This corpus represents the data collected within the Verbmobil project, which had as its task domain the negotiation of meetings and trip planning between two speakers.

The corpus consists of 1,658 spontaneous dialogs with 13,890 turns produced by 655 speakers. The text data comprises about 350K words, with a vocabulary of 6,680 unique words. The total duration of the recorded speech is 33:51:42 h.

Three kinds of filled pauses (*äh*, *ähm*, *hm*) and hesitation were regarded as the same event (filled pause) representing the same linguistic function. For detailed information about the nonverbals present in the corpus see Table 1.

In this paper we consider only nonverbal events; nonspeech sounds like paper rustle or squeak were eliminated from the textual data. All experiments were conducted using 10-fold speaker independent cross-validation.

3 Approach to Modeling of Nonverbals within Conversational Speech

As the baseline model we utilized a backoff trigram trained and tested using the SRILM toolkit [9]. In this study we evaluated four modeling approaches:

1. **Clean model:** Nonverbals are excluded both from data and language model.

2. **Regular model:** Nonverbals are modeled as regular words. Inclusion of nonverbal markers into the test set causes an issue: increased number of tokens to be predicted makes the direct perplexity comparison with the previous model inconsistent. In order to overcome this issue we calculated two values for this kind of model: ‘all tokens’ and ‘verbal only’. The former was calculated as usual on the full test set. For the latter we took into account only probabilities of the verbal tokens, thereby we estimate this value on the same amount of data as the perplexity value obtained with the ‘clean’ model.

3. **Omitted event model:** Nonverbals are modeled, but omitted from context. In this model we assume that the fluent context is a better predictor for the following word, thus the nonverbal token is excluded from the N-gram context and the preceding word is included instead. For instance, for the sentence with medial filled pause *ähm*:

   `Ja wunderbar <ähm> wiederhören`

   the following trigrams are generated:

   `Ja wunderbar <ähm>`   `Ja wunderbar wiederhören`

   but not the following trigram:

   `wunderbar <ähm> wiederhören`

   The nonverbal itself is considered as a usual word conditioned on the corresponding cleaned-up context. This is the ‘cleanup model’ in [4]. When evaluated only on the verbal tokens, this model becomes equivalent to the ‘clean’ model; therefore, only results including nonverbals are presented.

4. **Hidden event model:** modeling of nonverbals as hidden events. In this case speech is considered as a word stream with probabilistic events between words, some of which are hidden from direct observation, but are included in predicting context. The language model is trained on data containing the nonverbal tokens, but tested on data containing only verbal tokens. This approach is used for modeling of disfluencies and sentence or topic boundaries.

4 Experiments and Results

Our experiments show that including nonverbal tokens lowers the perplexity of the full test data (see Table 2). At the same time, including nonverbals into the model as regular words increases the perplexity of the verbal tokens. However, experiments including individual nonverbal types let us see that their effects differ. The overall gain is mostly due to the low local perplexity of the most frequent nonverbals, namely filled pauses, breath, and lip smack. For these event types the perplexity of the verbal tokens is higher than for the full test data. The difference is proportional to the frequency of the nonverbal type.

Table 2. Perplexity of the test data including different nonverbals. Highlighted values show an improvement compared with the baseline model with removed nonverbals.

Configuration	**Perplexity**			
Excluding all	66.85			
Including:	as regular word		as omitted event	as hidden event
	all tokens	verbal only		
– filled pauses	67.20	68.28	67.28	**66.27**
– breath	**62.59**	70.15	**63.22**	**65.88**
– laughter	67.02	66.86	67.04	**66.83**
– verbal noise	67.82	67.36	67.58	**66.74**
– throat clean	66.97	66.87	66.96	66.85
– swallow	67.08	66.89	67.06	**66.84**
– lip smack	67.07	67.93	67.06	**66.54**
Including all	**63.80**	73.88	**65.01**	**66.65**

As seen in Table 2, including any event type as a word into the language model actually slightly increases the perplexity, except for breath, which has a remarkable influence on the overall result. Breath is unlike the other nonverbals we consider here. Including it into the model as a regular word gives a major reduction in perplexity, while modeling it as an unobserved item reduces the beneficial effect.

The special status of breath events can be explained by their nature: breathing tends to occur at the end of full phrases within long sentences. The following sentence illustrates this fact:

`Ja und dann bin ich im Dezember ab siebzehnten weg <BREATH> bis Silvester dann.` (*Yeah so I will be away in December from the 17th <BREATH> see you on the New Year's Eve then.*)

We also found that breaths often occur before or next to the other nonverbals, especially filled pauses or lip smacks. Other modeling approaches for breath are the hidden event model or omitting them from context. However, experimental results show that the lowest perplexity is obtained with breath as a regular event within N-grams (see Tables 2, 3).

Table 3. Local perplexity at breath markers

Model	**Breath**		**Breath+1**		**Breath+2**	
	bigram	trigram	bigram	trigram	bigram	trigram
Regular	13.84	14.04	131.54	110.70	59.66	51.15
Omitted event	15.70	16.28	148.74	150.72	59.44	54.49
Hidden event	–	–	172.43	170.76	56.67	49.62
Clean	–	–	178.87	183.18	57.03	52.46

Filled pauses (FP) are of interest due to their importance for natural language understanding and their role as markers of linguistic and prosodic boundaries. Results presented in Table 4 confirm that FPs are better predictors for the following words and omitting them from context makes local perplexity significantly worse. At the same time, including filled pauses as words in the model does not improve the overall perplexity, but actually increases it (cf. Table 2).

Table 4. Local perplexity at filled pause positions

Model	FP		FP+1		FP+2	
	bigram	trigram	bigram	trigram	bigram	trigram
Regular	35.21	36.98	299.87	280.37	84.11	76.69
Omitted event	36.68	38.80	355.37	370.07	83.26	78.92
Hidden event	–	–	339.38	351.38	81.63	75.41
Clean	–	–	353.26	369.42	81.72	77.56

Previous research has shown that cutting context before medial filled pauses (using bigrams instead of trigrams) can improve language models on acoustically segmented sentences [5]. This can be explained by the role of medial FPs as linguistic segment boundaries, thus previous context often belongs to another phrase. However, our experiments on Verbmobil show that regular trigrams still provide lower local perplexity after FPs than regular bigrams. In order to check this we conducted an additional experiment just with those filled pauses which appear in the middle of sentences. The results are presented in Table 5.

Table 5. Local perplexity at medial filled pauses

Model	FP-M		FP-M+1		FP-M+2	
	bigram	trigram	bigram	trigram	bigram	trigram
Regular	41.70	44.16	337.43	319.59	88.41	80.35
Omitted event	43.19	46.08	394.28	413.20	88.39	84.98
Hidden event	–	–	380.44	392.36	86.51	80.74
Clean	–	–	393.98	414.48	86.75	83.46

As seen in Table 5, local perplexity after medial FPs in general corresponds to the values for all FPs (Table 4). An interesting result is that bigrams provide lower perplexity for the filled pauses themselves. Also, when FPs are removed from context (in omit and 'clean' models), bigrams predict following words slightly better than trigrams.

Experiments with Speech Recognition. A certain correlation exists between language model perplexity and word error rate (WER) [10]. However, even substantial reductions in perplexity do not always provide a marked benefit in word accuracy. To check to what extent breath markers can benefit recognition performance, we performed several experiments with speech recognition. The LM setup for speech recognition was slightly different from that for perplexity evaluation. We included into the model only those nonverbals which are most relevant for natural language processing and understanding, namely, filled pauses and laughter. Other nonverbal and nonspeech sounds were modeled acoustically, but with no output into the final recognition hypothesis ('invisible' event), so as to avoid insertions caused by misrecognition of noises as short words. In order to evaluate the role of breath specifically as a linguistic boundary marker, we removed initial and final breath markers in both hypotheses and references (i.e., they did not affect word accuracy).

Speech recognition was performed using the HTK toolkit [11]. For acoustic modeling we employed three-state left-to-right hidden Markov models/Gaussian mixture models (HMM/GMMs) with 32 mixtures. Sixteen nonverbal and nonspeech sounds were modeled with nine-state GMMs due to their longer average duration. In contrast to language models, we created separate acoustic models for three types of filled pauses (*äh*, *ähm*, *hm*) and hesitation to account for the different acoustic realizations. The feature set consisted of the typical 13 Mel-frequency cepstral coefficients (MFCC) including the 0th coefficient (energy) and delta and acceleration coefficients. Features were extracted from frames of 25 ms length sampled every 10 ms.

Table 6. Perplexity and word error rate for different language models

Model	Perplexity	WER	Nonverbal WER
Trigram excluding breath	67.38	30.21	47.89 †
Breath as hidden event	66.39	30.17	45.22 †
Breath as omitted event	65.54	29.03	42.23
Breath as regular word	64.16	28.97	42.43
Breath as word, eliminated	–	27.72	30.31 †

† *For these setups breath markers were excluded from both output and reference labels, thus the values were obtained just on two nonverbal tokens: filled pauses and laugh.*

Results presented in Table 6 show a 4.3% relative reduction of the word error rate in comparison to the setup where breath is modeled only as an acoustic event. Since breath markers convey little linguistic meaning, they can be removed from the output using post-processing. In this case the relative WER improvement increases to 8.2%, in spite of the higher perplexity of the verbal tokens alone for this language model type (see Table 2). Modeling breath as a hidden event allows us to apply the model on unmarked data, possibly from other corpora. In this case we also see a small improvement in both perplexity and word accuracy.

Analysis of the recognition rates of nonverbals shows high insertion rates for those nonverbals. Moreover, breath is inserted much more often than laughter or filled pauses. For example, when modeling breath as a regular word, relative to a total token count of 24,284, 21,203 are recognized correctly, and 7,223 are inserted incorrectly. When breath is excluded from the model, the proportions are roughly the same: 8,436 tokens in total, 6,797 are recognized correctly, with 2,401 insertions. However, when modeling breath as a regular word with eventual removal from the hypothesis, the number of insertions falls dramatically: 6,721 correctly recognized, with only 842 insertions. We surmise that breath is often substituted for short noises and verbal sounds, which in other setups would be recognized as filled pauses. Thus, breath can 'eat up' a substantial number of false insertions of other nonverbals; eliminating it from the output thus removes all these errors.

5 Conclusions and Future Work

Previous research showed that nonverbal sounds have rather strong local effect and do not influence the overall perplexity and speech performance result much [4,5,6]. Our experiments show that despite an evident gain in local predictions, increased perplexity of the verbal content almost nullifies this advantage or even leads to higher total perplexity. There is also a remarkable difference between the perplexities of the verbal tokens alone and the full test data. The effect correlates with the frequency of a certain nonverbal in the corpus. We can assume there is a threshold (about 1.5% for Verbmobil) after which perplexity of the full test data improves compared to the value obtained on the verbal part alone. If the frequency of a nonverbal type relative to all tokens is greater than 4–5%, including it in the model as a regular word may lead to a substantial improvement in both perplexity and word error rate. For the Verbmobil corpus, breath is an example of such a nonverbal type.

Other than benefiting perplexity and accuracy, including nonverbal tokens into the language model can enrich transcriptions with paralinguistic information, which may be important for natural speech processing and understanding. The most significant events for this purpose are laughter and filled pauses. Our experiments on the Verbmobil database show that modeling laughter and filled pauses as regular words is preferable compared to omitting these events from context or modeling them as hidden events.

For future work, we would like to employ the language modeling techniques considered in this paper together with detection methods based on the acoustic signal (e.g., as presented in [12]). This approach could further improve spontaneous speech recognition using some of the same phenomena that have been considered troublesome in the past.

Acknowledgments. The authors acknowledge the support provided by the federal state Sachsen-Anhalt with the Graduiertenförderung (LGFG scholarship). We also acknowledge the German Research Foundation (DFG) for financing our computing cluster used for parts of this work.

References

1. Ostendorf, M., Shriberg, E., Stolcke, A.: Human language technology: Opportunities and challenges. In: Proceedings of ICASSP, vol. 5, pp. 949–952. IEEE (2005)
2. Campbell, N.: On the Use of NonVerbal Speech Sounds in Human Communication. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) Verbal and Nonverbal Commun. Behaviours. LNCS (LNAI), vol. 4775, pp. 117–128. Springer, Heidelberg (2007)
3. Schultz, T., Rogina, I.: Acoustic and language modeling of human and nonhuman noises for human-to-human spontaneous speech recognition. In: Proceedings of ICASSP, vol. 1, pp. 293–296. IEEE, Detroit (1995)

4. Stolcke, A., Shriberg, E.: Statistical language modeling for speech disfluencies. In: Proceedings of ICASSP, Atlanta, GA, pp. 405–408 (1996)
5. Siu, M., Ostendorf, M.: Modeling disfluencies in conversational speech. In: Proceedings of ICSLP, vol. 1, pp. 386–389. IEEE, Philadelphia (1996)
6. Siu, M., Ostendorf, M.: Variable N-grams and extensions for conversational speech language modeling. IEEE Transactions on Speech and Audio Processing 8, 63–75 (2000)
7. Shriberg, E.: Spontaneous speech: How people really talk and why engineers should care. In: Proceedings of EuroSpeech, pp. 1781–1784 (2005)
8. Burger, S., Weilhammer, K., Schiel, F., Tillmann, H.G.: Verbmobil Data Collection and Annotation. In: Verbmobil: Foundations of Speech-to-Speech Translation, pp. 537–549. Springer (2000)
9. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proceedings of ICSLP, vol. 2, pp. 901–904 (2002)
10. Dietrich, K., Peters, J.: Testing the correlation of word error rate and perplexity. Speech Communication 38, 19–28 (2002)
11. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book (for HTK Version 3.4). Cambridge University Press, Cambridge (2000)
12. Fukuda, T., Ichikawa, O., Nishimura, M.: Breath-detection-based Telephony Speech Phrasing. In: Proceedings of INTERSPEECH, pp. 2625–2628 (2011)