JOINTLY RECOGNIZING MULTI-SPEAKER CONVERSATIONS

Gang Ji and *Jeff Bilmes*

Department of Electrical Engineering
University of Washington
Seattle, Washington, USA
{gang,bilmes}@ee.washington.edu

ABSTRACT

We suggest an approach to speech recognition where multiple sides of a conversation in a dialog or meeting are processed and decoded jointly rather than independently. We moreover introduce a practical implementation of this approach that demonstrates both language model perplexity and speech recognition word error rate improvements in conversational telephone speech. Specifically, we show that such benefits can be had if a n-gram language model, in addition to conditioning on immediately preceding words in an utterance, is also allowed to condition on the estimated dialog-act of the immediately preceding utterance of an alternate speaker.

***Index Terms*—** Speech recognition, multi-speaker, graphical models

1. INTRODUCTION

Humans are social beings and interact with and influence each other in subtle, complex, and diverse ways. At the phonetic level, individuals speaking in a dialog or group may after a time converge on an implicitly agreed upon rate-of-speech, pitch range, style of turn timing, loudness quality, or other prosodic pattern [1, 2]. At the linguistic level, is has been argued that conversations are, on balance, "easier" to ascertain than monologues because of an alignment process that eventually occurs between participants in order to allow each participant to better and more efficiently meet their underlying goals of a discourse [3, 4, 5]. And at more abstract levels still, the outward behavior of an individual, including gestures and other mannerisms (such as gait), will adjust depending on the set of individuals who are participating in an interaction [6]. This property, moreover, has recently been observed in computational social-network analyses of measured situated speech data [7].

Indeed, most large vocabulary continuous speech recognition (LVCSR) systems operate on corpora that originated from a conversation amongst two [8] or more [9] individuals, i.e., where the aforementioned automatic and eventual alignment process almost certainly occurs. LVCSR systems have as their basis statistical machine learning and information theory, where representing such interactions can be shown mathematically to have benefit. For example, it has long been known [10] that when two random sources X_1 and X_2 are not independent (i.e., $p(X_1, X_2) \neq p(X_1)p(X_2)$), one may obtain a more efficient code for the two if they are encoded jointly rather than separately. In our case, X_1, X_2 might be two successive utterances spoken by two different speakers in a dialog, or they might correspond to the entirety of two sides of a conversation. Similarly, in the domain of pattern classification where class labels Y_1, Y_2 exist corresponding to the correlated sources X_1, X_2, there is potential for improved accuracy if Y_1, Y_2 are decided jointly based on the information in both X_1, X_2, rather than Y_i being decided based only on X_i, for $i = 1, 2$.

Most LVCSR systems, however, operate on data that has been excised from its original context, where speech waveforms arising from an interactive dialog between multiple individuals are first segmented into isolated single-person utterances. These utterances are then typically processed by the recognizer entirely independently of each other. When this is the case, any resulting statistical dependency between interlocutors is lost entirely. It is perhaps unfortunate that the very interactions and influence that social scientists study is that which LVCSR systems ignore, as modeling such influence could greatly benefit the practice of speech recognition, not to mention the field of machine translation which might similarly yield benefits.

Of course, representing such phenomena has its downside as well, as the curse of dimensionality will place greater requirements on the amount of training-data, to produce such a model, and on the computational resources (memory and available computing power) to decode with such a model. A research goal, therefore, is to develop new practical multi-party conversational decoding strategies that balance the trade-off between model accuracy (by representing inter-speaker dependency), and practical realism (any such method must consider the above limitations).

Fortunately, there is nowadays wide availability of large amounts of training data that are sufficient to demonstrate the utility of such an approach. Even so, past work has occurred only infrequently on inter-speaker dependency modeling for speech recognition. As best as we can tell, only one study did this [11] where the authors represented inter-speaker dependency at the language level by allowing the language model of a speaker's first words in an utterance to condition on the most recently spoken word of a previous speaker — the study showed language model perplexity benefits in both conversational telephone speech and meeting scenarios, but no word-error results were given. In this work, we propose a method where the language model of all the speaker's words in an utterance may condition on the most recently spoken dialog act [12] of a previous and alternate speaker. By representing this dependency, we show that we achieve an improvement in both language model perplexity and speech recognition word error on the conversational corpora Switchboard.

2. MULTI-SPEAKER CONVERSATION MODELING

In standard LVCSR, we are given an isolated length-T acoustic segment of speech $\bar{x}_{1:T}$ (typically in the form of a list of feature vectors for a single acoustic channel) and the goal is to determine the resulting length-N sequence of words $w_{1:N}$ where both the sequence and its length N is unknown. This is typically done using a (generative) hidden Markov model (HMM) to produce a joint model $p(\bar{x}_{1:T}, w_{1:N}) = p(\bar{x}_{1:T}|w_{1:N})p(w_{1:N})$, where $p(\bar{x}_{1:T}|w_{1:N})$ is the acoustic model, and $p(w_{1:N})$ is the (often n-gram) language model, and then decoding occurs via $w^*_{1:N^*} \in \operatorname{argmax}_{N,w_{1:N}} p(\bar{x}_{1:T}|w_{1:N})p(w_{1:N})$.

We propose instead a general framework of multi-speaker conversation modeling. We have separate acoustic feature vectors (say from different microphones) for K audio channels corresponding to K speakers $\bar{x}^1_{1:T}, \ldots, \bar{x}^K_{1:T}$. We assume in this work that the number of speakers is known a priori and each speaker has her own channel. We also assume that all acoustic sequences are of the same length (meaning that much of each speaker's acoustics might represent silence, although an implementation might choose to optimize this). We wish to decode the words for each speaker $w^1_{1:N_1}, \ldots, w^K_{1:N_K}$, where $w^k_{1:N_k}$ is a length-N_k *word sequence* corresponding to speaker k. The length of $(x^k_{1:T}, w^k_{1:N_k})$ might span in total over an entire discourse, rather than over a single utterance, so the sequence lengths will be much longer than is typically the case for isolated utterances. Moreover, each speaker will speak a different number of words (i.e., N_i is typically unequal to N_j for $i \neq j$). We assume the existence of a model of the multi-speaker joint distribution $p(\bar{x}^1_{1:T}, \ldots, \bar{x}^K_{1:T}, w^1_{1:N_1}, \ldots, w^K_{1:N_k})$ and we wish to decode via:

$$(w^{*1}_{1:N_1^*}, \ldots, w^{*K}_{1:N_K^*}) \qquad (1)$$
$$\in \operatorname*{argmax}_{N_{1:K}, w^1_{1:N_1}, \ldots, w^K_{1:N_k}} p(\bar{x}^1_{1:T}, \ldots, \bar{x}^K_{1:T}, w^1_{1:N_1}, \ldots, w^K_{1:N_K})$$

Such a joint distribution would allow for the representation of a variety of linguistic inter-speaker phenomena. For example, phonetic entrainment [1] would mean a dependency would exist directly between $\bar{x}^i$ and $\bar{x}^j$ $(i \neq j)$. Linguistic alignment [4] would imply dependency between the corresponding word strings. The implementation details of such dependencies will determine both the model tractability and any demands on amounts of training data. Indeed, there is a vast diversity of modeling choices implicit in Equation 1.

The work of [11] fits entirely within the above framework. Therein, K was set to 2 (in the meeting scenario all other speakers were collapsed down to a single individual for the sake of dependency modeling), and a standard tri-gram language model for each speaker $p(w_t|w_{h(t)})$ was augmented with information about the most recent word spoken by the alternate speaker as in $p(w_t|w_{h(t)}, a_t)$ where $w_{h(t)}$ is the history up to time $t-1$ and a_t is the most recent (relative to t) alternate speaker's word. The authors demonstrated improved perplexity on both conversational telephone and meeting data.

There are several potential problems, however, with this implementation. The first is that this language model was used only for the first few words of the current speaker's utterance, the rest of the utterance used a language model independent of the other speaker (the authors found no benefit to having a dependency between a_t and words w_τ for $\tau > t+1$). The second is that the variable a_t may have any value in the lexicon, so in the 3-gram case, the model $P(w_t|w_{t-1}, w_{t-2}, w_a)$ has as many data demands as does a four-gram language model.

In this work, we consider a language model that applies to all words in the current utterance (addressing problem 2 above) and that conditions on only the dialog act (DA) [12] of the alternative speaker (addressing problem 1). Discourse patterns in natural conversations and meetings are well known to provide useful information about human conversational behavior. DAs, which reflect the functions that utterances serve in a discourse, may therefore be beneficial as a knowledge source, even across speaker. In our work, a 3-gram language model therefore becomes:

$$P(w_t|h_t) = P(w_t|w_{t-1}, w_{t-2}, d_t) \qquad (2)$$

where d_t is the dialog act (estimate) of the previous sentence of the alternative speaker relative to word w_t. This model, moreover, is applied to all words of a current utterance rather than just the first.

3. EXPERIMENTS

There are many ways to implement the aforementioned ideas in a practical automatic speech recognition system. Ideally, we would decode the K channels simultaneously. Even single-channel ($K = 1$) speech recognition decoding is expensive, however, so typically a multi-pass strategy is used where multiple highly-scoring hypotheses of a first pass system are represented using a lattice which is then used to constrain the search in latter passes which rescore the first-pass lattice hypotheses. Extending this to $K > 1$ channels, each channel should have its own lattice, and these multiple lattices should somehow be jointly decoded, using the above language model as a means through which hypotheses in different lattices may interact.

In [13], it was shown how graphical models (GMs) can be used to represent a lattice — more recent work [14, 15] shows how graphical models can express a system that corresponds to more than one lattice and how the standard GM decoding algorithms can automatically produce the joint alignment of multiple lattices. Such an approach can be used here as well, where each lattice has its own decoding sub-graphical model which consists a rescoring-word node that depends both on the two previous word nodes (in the tri-gram case) as well as a DA node for the previous speaker. These subgraphs can then be combined with a separate GM that represents a DA tagger [16]. All coupled together, such a graph would represent the multi-speaker speech recognition and dialog act tagging of a multi-channel dialog or meeting. Such a modular approach is certainly attractive from a conceptual perspective, since pre-existing and pre-trained GM components can be glued together in an almost unmodified form to produce one large joint decoder. On the other hand, the "product" of two or more channels, coupled with a simultaneous DA tagger for each channel would produce an enormous state space, even with each channel's state space reduced via a lattice representation.

Another approach, therefore, that is more tractable but is an approximation of the above, is to perform the lattice rescoring process in two steps. The first step produces speech lattices for each channel, like the above, but then a separate DA tagger is used to estimate dialog act for each utterance based either on the one-best lattice hypotheses (what we adopt herein), or even on the entire lattice itself. After this is done, we can use the DA tag hypothesized from the first-pass system as an observed variable in a second-pass system to represent the inter-speaker dependency. Note, the DA tag doesn't itself entail such dependency, rather the DA tag is used to condition on in a DA-based language model (Equation 2) that considers the dependency of utterance words on the dialog act of the most recent sentence from another speaker. In such a case, the decoding system considers only one channel (and in fact one utterance) at a time, and therefore the complexity is significantly less that of the first approach, and in fact is not much more difficult than a standard ASR system. The primary change is in the language model.

In this work, we have in fact attempted both of these approaches, but found that the first case required more memory for state-space

Table 1. Switchboard Dialog Act tag accuracy using GMTK

model	dev set	test set
generative model	85.1	83.0%
hidden backoff model	86.7	84.2%

Table 2. Perplexity results using dialog act based language model

model	dev set	test set
trigram	335.3	381.7
DA-based trigram	302.0	344.5
improvement	14.1%	9.7%

representation than what was available to us at the time (8GB). However, the second method is quite practical and produces good results (see below).

For either approach, we need access to a good dialog act tagger, a problem that has many possible solutions. In this work, we have adopted the DA tagger models described in [16, 17]. The training set for our dialog act tagger was taken from the results of the 1998 JHU workshop [18], which consists 1155 Switchboard phone conversations with 186467 sentences. For the tagged data in [18], we extracted three fifth (sw00utt–sw10utt) as the training set, one fifth (sw11utt–sw12utt) as the development set, and the rest one fifth (sw13utt) as the test set. Similar to [16], we have used the graphical models toolkit (GMTK) [19] to train and test the dialog act tagger. We tested two models: the generative model [16] and the "hidden backoff model" as described in [17]. The primary novel feature of the hidden backoff model is that sub-DA states are assumed to exist but are not labeled in the training data. Using a hidden state variable, these hidden states are learned in a backoff-model via the use of an embedded-Viterbi style EM algorithm — See [17] for details. The results are shown in Table 1, and are consistent with [17] in stating that the hidden-backoff approach produces a high-quality DA tagger.

Before proceeding to our speech recognition task, we wanted to ensure that there were at least perplexity gains for our DA-based language model (Equation 2) over a baseline model. We used the hidden backoff model trained in the previous step as the dialog act tagger and tagged the rest of the Switchboard corpus and the entire Fisher corpus. The overall data set has 1991 conversations and 305116 sentences. Based on this, we trained a DA-based trigram language model $P(w_i|w_{i-1}, w_{i-2}, d_a)$ using a factored language model (FLM) [20] as in the SRI language model toolkit [21]. In an FLM, the model may contain different backoff orders, different smoothing techniques and so on. These parameters were optimized using genetic search on a development set consisting of the eval1998 and eval2000 data sets which contain 40 conversations with 3586 sentences. The best obtained model has as backoff path [20] in $P(w_t|w_{t-1}, w_{t-2}, d_a)$ as follows: first drop side DA information d_a, then w_{t-2}, and w_{t-1} to a unigram, and finally drop w_t to a uniform distribution. Each level uses Kneser-Ney smoothing. We used this model to calculate perplexity on the test set, which is eval2001 which contains 60 conversations with 5859 sentences. The results are shown in Table 2. Quite promisingly, the table shows that the DA-based trigram model can indeed help reduce perplexity, which means that knowing the DA of the alternate speaker, at the very least, makes one more confident about the current correct set of words.

We tested our trained DA-based trigram FLM on the switchboard eval2001 test set, where thin word lattices were generated by the SRI MFCC-based within-word system[22]. Word lattices were represented using dynamic Bayesian networks as described in [13]. The decoding graphical model used in our experiment is shown in Figure 1 — limited space precludes us from fully describing this model, but we give an outline herein. At the high level, the graph may be seen as having two parts: the upper part is a graphical models representation of word lattices, as described in [13]. The lower part is the rescoring model. A middle part of the graph contains a variable (`consistency`) that glues the lattice together with a new rescoring language model. This variable ensures that only predictions from the word lattice are matched for trigram rescoring. The bottom part of the graph includes an observed variables (`side DA`) that contain the most recent DA of the alternate speaker. That is, the probability of `rescore word` may depend on two previous word tokens as well as the DA.

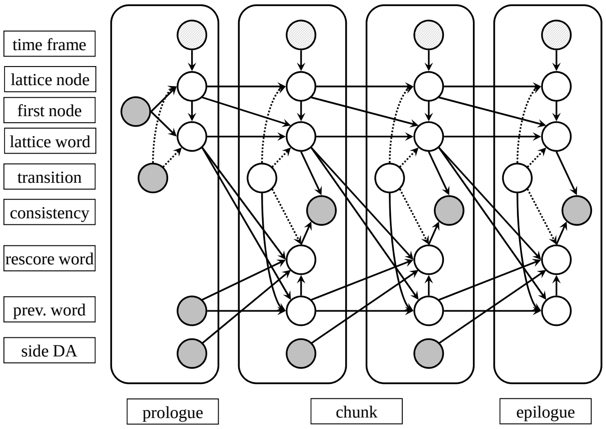

Fig. 1. Decoding graph using a word lattice and a DA-based trigram

Table 3. Word error rate using dialog based language model

model	dev set	test set
bi-gram baseline	29.3%	28.5%
tri-gram GMTK rescore	27.8%	26.0%
DA-lm	27.0%	25.5%

We tested the DA-based trigram model trained as described above. In order to produce DA tags, the one-best hypothesis from the lattice is used as input to the DA tagger, and the resulting tags are considered as "truth", even though they might have errors. We note again that the DA-dependent language model was trained on tags possibly containing such errors (we did not evaluate oracle DA tags as that would be less realistic). The DA-based LM is compared with a normal tri-gram rescoring LM that has been trained on the same data (Switchboard and Fisher) but without the use of DA information.

All experiments, again, were tested using GMTK and the results are given in Table 3. The "baseline" results row shows the word-error (WER) of the one-best results only from the word lattices. The "GMTK rescore" WER shows the word error rates of standard trigram rescoring from the lattices using GMTK. The last row, "DA-lm", shows our new model where prediction is based on also having access to the previous alternate speaker's DA tag. The results show, both on the development set and the test set, a small but significant improvement in WER over the trigram model. Informally, we have found that the nature of the errors that are corrected are quite different than what might be expected when going to a 4-gram language model. For example, if the previous DA is a "question", then that significantly changes the set of likely responses relative to the case where the DA is a "back channel". Ideally, however, future work will lead to detailed error analysis being performed to enable the full study of the nature of the errors that are corrected, an activity that might itself also lead to either novel linguistic insight or confirmation of existing theories of discourse patterns.

4. CONCLUSIONS

We described a new model for speech recognition whereby multiple speakers in a conversation are decoded jointly, rather than separately, thus taking advantage of any dependency across speakers. We have produced an initial and practical implementation of this idea: we have augmented a standard language model with an estimate of the dialog act of a sentences spoken by other speakers in a conversation. Our new model yields improvements in both language model perplexity and speech recognition word error.

In future work, we wish to make progress on full simultaneous decoding, e.g., the first approach mentioned in Section 3. Recent work showing how different segmentations can yield benefits [23] (albeit for translation) supports the case that this would yield further improvements, as a fully joint multi-speaker decoder would, in a sense, be deciding both the segmentation and word strings for multiple speakers simultaneously and jointly.

5. REFERENCES

[1] John Local, "Variable domains and variable relevance: interpreting phonetic exponents," *Journal of Phonetics*, vol. 31, no. 3-4, pp. 321 – 339, 2003, Temporal Integration in the Perception of Speech.

[2] John Local, "Phonetics and talk-in-interaction," in *International Congress on Phonetic Science (ICPHS)*, Barcelona, 2003.

[3] Susan Brennan and Herbert Clark, "Conceptual pacts and lexical choice in conversation," *J.Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, no. 6, pp. 1482–1493, 1996.

[4] Martin Pickering and Simon Garrod, "Toward a mechanistic psychology of dialogue," *BEHAVIORAL AND BRAIN SCIENCES*, vol. 27, pp. 169–226, 2004.

[5] Simon Garrod and Martin Pickering, "Why is conversation so easy?," *TRENDS in Cognitive Sciences*, vol. 8, no. 1, pp. 8–11, Jan 2004.

[6] A. Dijksterhuis and J.A. Bargh, "The perceptionbehavior expressway: automatic effects of social perception on social behavior," *Advances in Experimental Social Psychology*, vol. 33, pp. 1–40, 2001.

[7] D. Wyatt, T. Choudhury, J. Bilmes, and J. Kitts, "Towards the automated social analysis of situated speech data," in *Proceedings of UbiComp*, Seoul, S. Korea, September 2008.

[8] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP-1992*, 1992, pp. 517–520.

[9] N. Mirghafori, A. Stolcke, C. Wooter, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf, "From switchboard to meetings: Development of the 2004 ICSI-SRI-UW meeting recognition system," in *Proc. ICASP*, Jeju, Korea, October 2004.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, Inc., 1991.

[11] Gang Ji and Jeff Bilmes, "Multi-speaker language modeling," in *Proc. HLT/NACAC-2004*, Daniel Marcu Susan Dumais and Salim Roukos, Eds., Boston, Massachusetts, USA, May 2 - May 7 2004, pp. 133–136, Association for Computational Linguistics.

[12] J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, 1969.

[13] Gang Ji, Jeff Bilmes, Jeff Michels, Katrin Kirchhoff, and Chris Manning, "Graphical model representations of word lattices," in *IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT2006)*, Palm Beach, Aruba, December 2006.

[14] Hui Lin, Alex Stupakov, and Jeff Bilmes, "Improving multi-lattice-alignment based spoken keyword spotting," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009.

[15] Hui Lin, Alex Stupakov, and Jeff Bilmes, "Spoken keyword spotting via multi-lattice alignment," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brisbane, Australia, September 2008.

[16] Gang Ji and Jeff Bilmes, "Dialog act tagging using graphical models," in *Proc. ICASSP-2005*, Philadelphia, PA, March 2005.

[17] Gang Ji and Jeff Bilmes, "Backoff model training using partially observed data: Application to dialog act tagging," in *Proc. HLT/NACAC-2006*, New York, NY, June 2006.

[18] Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, and Carol Van Ess-Dykema, "Dialog act modeling for conversational speech," in *Proceedings of the AAAI Spring Symposium on Applied Machine Learning to Discourse Processing*, 1998, pp. 98–105.

[19] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," in *ICASSP-2002*, 2002, vol. 4, pp. 3916–3919.

[20] Jeff Bilmes and Katrin Kirchhoff, "Factored language models and generalized parallel backoff," in *Proc. of HLT/NACACL-2003*, 2003, pp. 4–6.

[21] Andreas Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICASSP-2002*, Denver, CO, September 2002, vol. 2, pp. 901–904.

[22] A. Stolcke, I.Bulyko, B. Chen, H. Franco, V.R. Gadde, M. Graciarena, N. Morgan, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, "Development of the 2004 SRI/ICSI/UW speech to text system," in *Proceedings of DARPA 2004 Rich Transcription Workshop*, 2004.

[23] Sharath Rao, Ian Lane, and Tanja Schultz, "Optimizing sentence segmentation for spoken language translation," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium, September 2007.