CORPUS-INDEPENDENT HISTORY COMPRESSION FOR STOCHASTIC TURN-TAKING MODELS

Kornel Laskowski

Carnegie Mellon University
Pittsburgh PA, USA

Elizabeth Shriberg

Microsoft Speech Labs
Mountain View CA, USA

ABSTRACT

Stochastic turn-taking models use a truncated representation of past speech activity to specify how likely a speaker is to talk at the next instant. An unanswered question in such modeling is how far back to extend the conditioning context. We study this question using Switchboard (English, telephone) and Spontal (Swedish, face-to-face) conversations. We also explore whether to trade off precision with range when moving backward in the history. We find that (1) a nearly logarithmic compression of history is optimal, for both speaker and interlocutor; (2) the absolute duration of the conditioning context is at least 7 seconds; and (3) the compression scheme generalizes remarkably well across the two different corpora.

Index Terms— Turn-taking, conversational speech, diarization, dialogue, speech activity.

1. INTRODUCTION

Stochastic turn-taking models use statistics, as opposed to rules, to predict when speakers talk versus not talk in conversation. The approach has roots in early studies [1, 2, 3] that conditioned predictions on information about only past speech activity (no words or other linguistic information), and showed that turn-taking can be viewed as a Markovian process. Initial work looked only at the speech activity patterns of a single talker [1]; later work expanded the modeling to include an interlocutor [2, 3].

Such models use only binary speech activity as features. They exclude words, prosody, and gestural or other visual cues. In this regard they are obviously too simplistic to reflect the complex process of human conversation. They also suffer from other disadvantages, notably practical limitations on model complexity given available data to train the model. Nevertheless, the models offer a number of advantages for the study of conversation and for real-time modeling of incipient speech in conversational speech applications. First, even though only speech activity is used, these features correlate with discourse-relevant information. For example, very short regions of speech between pauses correlate with backchannels such as "uh-huh". Speech activity features are both simple to compute and privacy-sensitive. Furthermore, they can be applied across corpora in which conversations may differ in topic, type of interaction, and choice of language. Finally, the models can be used as prior knowledge to constrain search in nearfield speech activity detection systems, in farfield speaker diarization systems, and in real-time dialogue systems for low-latency prediction.

Recent years have seen renewed interest in such models [4, 5, 6]. Yet an important question that remains unanswered is just how far back to look. It appears that whenever increasingly older context is added, models either continue to improve, or become intractable due to lack of sufficient data to train higher-order models [6].

The goal of this paper is to address the question of how far back to extend the history, to best predict upcoming speech. In asking this question, we also consider the possibility that the precise temporal location of conditioning events becomes fuzzy as those events recede into the past. Should that turn out to be true, the sensitivity of a model to variations in timing would not be uniform but would decay monotonically, making it possible to obtain greater range at the expense of precision. These concerns have implications for human processing models, particularly for models of attention allocation. They may also have impact in speech applications, for example in the real-time allocation of resources in dialogue systems and in the design of features that capture lexical or prosodic information.

Our investigation is carried out using a fixed modeling framework and two different corpora, described in Section 2. We intentionally chose corpora that differ in language, style, and presence of visual cues. Model performance with history compression is compared to that using standard history truncation in the experiments of Section 3, for both within-corpus and cross-corpus prediction. We discuss some implications and conclude in Section 4.

2. METHOD

2.1. Data and Segmentation

Our first corpus is the Switchboard-1 Corpus [7], as re-released in 1997. It consists of 2435 telephone conversations, each approximately 10 minutes in duration. We divide the data into three speaker-disjoint sets, such that our TRAINSET, DEVSET, and TESTSET consist of 762, 227, and 199 conversations, respectively. During this division process, it was not possible to allocate 1247 conversations because their two speakers had already been placed in different sets. Since we are modeling speaker state, reference speech/non-speech segmentations were used. These were obtained from the available forced alignments [8] for each conversation side.

To relate the study to previous work [6], we compare the Switchboard-1 Corpus with the Spontal Corpus [9]. The latter consists of 30-minute face-to-face conversations in Swedish, with each conversant participating in only one conversation. The Spontal training, development, and test sets consist of 23, 6, and 6 conversations, respectively, as in [6]. Only an automatically produced segmentation is available, produced as described in [10].

2.2. Stochastic Modeling

We use the modeling framework described in [6]. The segmentations for the two sides to each conversation are time-aligned and sampled at a non-overlapping frame step of 100 ms. This corresponds approximately to the duration of half of a syllable, and ensures that no

speech is discarded. The process results in a discrete $2 \times T$ binary-valued matrix, or *chronogram*, where T is the number of 100-ms frames in the conversation. A chronogram cell has the value ■ if the speaker represented by the cell's row spoke for the majority of the interval corresponding to the cell's column, and □ otherwise.

Chronograms are treated as vector-valued Markov processes, allowing us to factor their likelihood into a product of conditional n-gram likelihoods. Each such likelihood factor can be further decomposed into two factors, one corresponding to each of the two conversants. We build two types of model: (1) conditioning each participant's speech activity only on that participant's past speech activity, and (2) conditioning each participant's speech activity on the past speech activity of both participants. We refer to these as unconditionally independent (UI) and conditionally independent (CI) models, respectively.

Maximum likelihood estimates for all n-gram probabilities are inferred using training data, and are then smoothed recursively using linear interpolation with lower-order n-gram models. The smoothing, described in detail in [6], uses a global relevance parameter optimized on development data.

Models are assessed using the average negative log-likelihood over every cell, in each column of each row of each test set chronogram. This measure is identically the cross entropy rate of the Markov process assumed to produce the chronograms, given a trained model. Cross entropy rates in this work are expressed in bits per 100-ms frame; a value of unity would indicate that chronograms are completely random, whereas a value of zero that they are always □ or always ■. The difference between 1st-order UI and CI models is significant: in a speech activity detection experiment on the `rt05s_eval` and `rt06s_eval` data sets, modeling speakers as not independent reduced NIST error rates by 40% relative (cf. Figure 11.29 in [11]).

3. RESULTS

3.1. Standard Representation of History

We apply the modeling described above to the Switchboard-1 (Swb-1) data, constructing UI and CI models with $\tau \in [1, 10]$ frames of context. The results are shown in Figure 1, in red. Also shown, in blue, are curves obtained using the Spontal corpus as in [6].

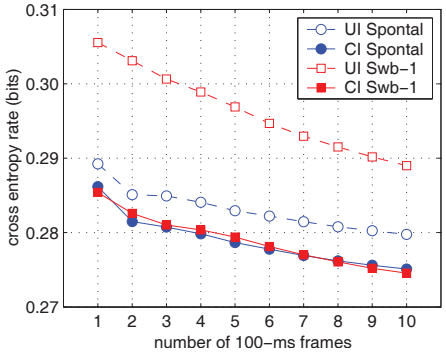
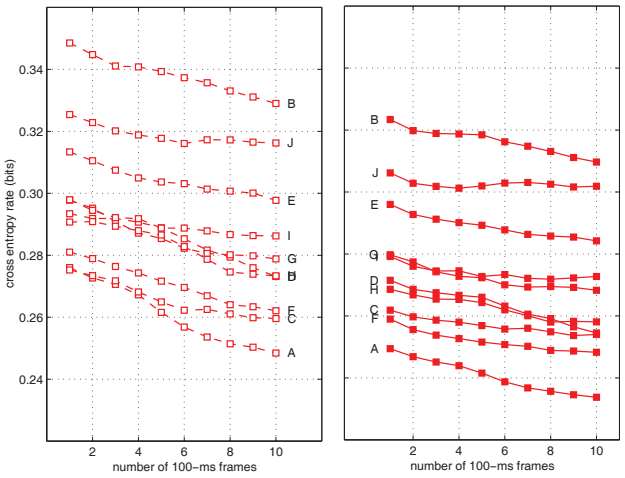


Fig. 1. Cross entropy rate as a function of the number of 100-ms frames in the conditioning context. UI: only target speaker in history, CI: target speaker plus interlocutor in history.

The main observation of interest in Figure 1 is that all curves exhibit monotonically decreasing cross entropy rates as the conditioning context grows by one 100-ms frame at a time. It appears that at 1 second, after 10 frames have been observed, the cross entropy rate continues to fall.

There are also some interesting differences between Switchboard-1 and Spontal conversations. For example, when the interlocutor is ignored (i.e., the UI condition), conversants' speech activity appears more difficult to predict in Switchboard-1 than in Spontal. This is true for any model order. Evidently, the past durations and sequencing of one's own talkspurts and gaps exhibit more variation in Switchboard-1 than in Spontal.

On the other hand, we believe that the nearly identical performance of the CI model for both corpora is largely a coincidence. We feel the shape of the curve is meaningful, but the absolute performance depends somewhat on the data. Individual speakers and speaker pairs have different overall rates of speech and overlap frames — a topic we are exploring in separate work. But to briefly illustrate this here, Figure 2 shows results for 10 randomly selected TESTSET conversations. It indicates that there is a wide band of variation around the mean shown in Figure 1, while retaining roughly parallel trends.



(a) UI model speaker-only history

(b) CI model speaker+interlocutor history

Fig. 2. Cross entropy rate as a function of the number of 100-ms frames in the conditioning context, for 10 randomly selected Switchboard-1 conversations, labeled "A" through "J".

We note that it occasionally happens that the CI curve for one conversation is higher than the UI curve for another. Also, the UI and CI curves are not decreasing with τ for all conversations; occasionally a given curve rises only to fall again for larger τ. However, it appears to always be the case (not shown) that for any given conversation, the CI curve lies below the UI curve.

3.2. History Compression

We now investigate an alternative to conditioning on a uniformly discretized past. Any proposed alternative involves a trade-off among 3 factors: (1) the absolute duration of the modeled context, (2) the granularity with which one tiles that context, and (3) the number of degrees of freedom of the model. In the current work, we keep the

model order fixed[1]. The question then becomes one of allocation of model resources to yield optimally predictive conditioning. Based on Figure 1, we expect that activity older than 1 second may help. To expose a model to such older events, we need to either skip earlier frames or to map frames in some manner onto larger windows.

In this paper, we choose to merge rather than skip frames, independently for the chronogram rows corresponding to the target speaker and the interlocutor. In mapping adjacent frame values to a single value, we employ majority voting. Ties for windows that subsume an even number of frames are resolved by assigning ■ if at least half of the subsumed frames are ■, and □ otherwise[2].

To determine the optimal geometry of mapping windows, we performed an automated search as described in Algorithm 1. Each evaluation of cross entropy rate involves training a model using TRAINSET, smoothing parameter optimization using DEVSET, and model scoring using TESTSET; only Switchboard-1 data was used.

Algorithm 1 Search for a windowing policy given 10 windows

current frame index: $\tau = 1$
number of found windows more recent than τ: $M = 0$
found window sequence $W = \emptyset$
while $M < 10$ **do**
  for $m = 1$ **to** ∞ **do**
    posit $10 - M$ windows of m frames at $\geq \tau$
    $\xi\,[m]$ = cross entropy rate
  end for
  optimal window size at τ: $m^* = \arg\max_m \xi\,[m]$
  for $i = 1$ **to** $10 - M$ **do**
    hypothesize i windows of size m^* at index τ
    consider the advance $\tau'\,[i] = \tau + i \cdot m^*$
    for $n = m^* + 1$ **to** ∞ **do**
      posit $10 - M - m^*$ windows of n frames at $\geq \tau'\,[i]$
      $\xi'\,[n]$ = cross entropy rate
    end for
    cross entropy rate at $\tau'\,[i]$: $\xi^*\,[i] = \max_n \xi'\,[n]$
  end for
  optimal number of windows at τ of size m^*: $i^* = \arg\max_i \xi^*\,[i]$
  place i^* windows of size m^* in W
  $\tau = \tau'\,[i^*]$
  $M = M - m^*$
end while

The algorithm starts at the instant immediately preceding the predicted frame, and considers tiling the history with 10 uniformly wide windows. After finding the best uniform window width for the history beyond the current instant τ, it considers retaining only a few windows of that width, windowing frames beyond those with a larger uniform width. It picks the window width m^* and number of frames i^* which yield the lowest cross entropy rate, increments τ with $m^* \cdot i^*$, and continues until all 10 windows are specified. As can be seen, a shrinking in window size as the algorithm proceeds is never entertained. However, nothing prevents the algorithm from settling on uniform windowing or on piecewise uniform windowing.

Applying Algorithm 1 to only the target conversant's history (UI model) led to a final sequence of windows with durations { 100 ms, 100 ms, 100 ms, 200 ms, 400 ms, 500 ms, 500 ms, 1000 ms, 1000 ms, 3300 ms }, from most recent to least recent. The total history duration is 7.2 seconds. Keeping this sequence fixed, Algorithm 1 was then applied to the interlocutor's history, relevant to CI models. The found sequence consists of windows with durations { 100 ms, 100 ms, 100 ms, 300 ms, 300 ms, 300 ms, 900 ms, 2000 ms, 2000 ms, 2000 ms }. The total resulting duration in the interlocutor row is 8.1 seconds. Both windowing policies, shown in Figure 3 for comparison, retain the maximum available precision of 1 frame for the first three windows, and then widen windows to achieve a nearly logarithmic taper which terminates at around 7.5 seconds.

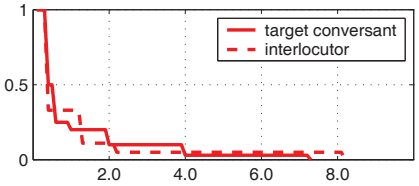


Fig. 3. Results of empirical stepwise search to find an optimal windowing policy given 10 windows for target conversant (blue) and interlocutor (red), as a function of time (along the x-axis) looking backwards. Height along the y-axis is the width (in number of frames) of the underlying policy window. Interlocutor policy search starts with best target speaker policy already in place.

3.3. Multi-Corpus Comparison

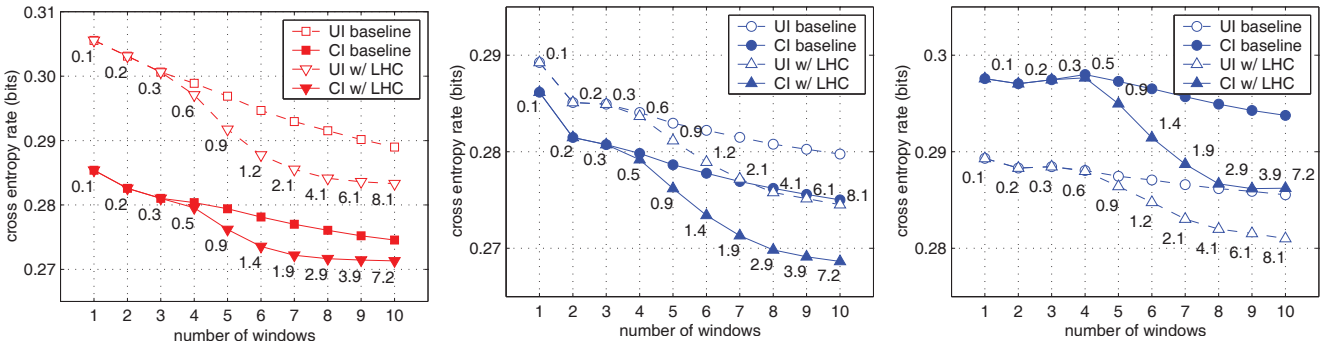
We now compare the effect of the optimized windowing policies on the curves shown in Figure 1, which use standard representations of history. The comparison for the Switchboard-1 Corpus is shown in Figure 4(a). As can be seen, both UI and CI models benefit from more conditioning history, at every model complexity explored except for 1, 2, and 3 degrees of freedom. This is of course because both optimized window sequences retain only a single frame in each of their first three windows, making the compressed and uncompressed schemes identical when only a small number of model parameters is available.

Since the window sequences were found by evaluating cross entropy rate on TESTSET, Figure 4(a) does not offer a truly independent assessment. It also doesn't let us know how well the policy generalizes to other types of data. To investigate these questions, we apply the window sequences found using Switchboard-1 data to the Spontal corpus, in Figure 4(b). Models are trained, tuned, and evaluated on Spontal data only, making the resulting curves comparable to those shown for Spontal in Figure 1. As for the Switchboard-1 curves in Figure 4(a), history compression is seen to be beneficial. In fact, a compressed-history UI model achieves parity with an uncompressed-history CI model, at the right of the figure.

Finally, we apply the window sequences as obtained using Switchboard-1 data, as well as the models trained and tuned to the Switchboard-1 TRAINSET and DEVSET, to the Spontal test data. The results are shown in Figure 4(c). We observe that the UI curve exhibits approximately the same profile as when models are trained on Spontal, but with gentler slope. Interestingly, the CI curve is *above* the UI curve. This tells us something about corpus differences: models of interlocutor behavior obtained using Switchboard-1 are actually not appropriate for Spontal data. What is relevant to the current work is that despite the reversal of curves due

[1]Fixing context size, to instead compare compressed versus uncompressed tiling, is complicated by issues of smoothing; the rightmost UI and CI models in Figure 1 are already an 11-gram and a 21-gram, respectively.

[2]We allow ■ to dominate □ because we treat speaking as the marked condition. In particular, conversants appear to generally avoid speaking simultaneously but do not avoid not-speaking simultaneously.



(a) train: Swb-1, test: Swb-1 (b) train: Spontal, test: Spontal (c) train: Swb-1, test: Spontal

Fig. 4. Effect of logarithmic history compression (LHC) on smoothed n-gram model performance. Number of mapped windows along the x-axes, cross entropy rate in bits along the y-axes. Points along LHC curves annotated with total history duration, in seconds. UI: only target speaker in history, CI: target speaker plus interlocutor in history. Curve color corresponds to that of the test set in Figure 1.

to training n-gram counts on different-domain data, history compression determined using different-domain data leads to consistent improvements, for both the UI and CI curves.

4. DISCUSSION & CONCLUSIONS

The results of the previous section were obtained using speech activity features, and relate to the prediction of speech activity using those same features. This makes sense: a real-time task may find features that take more time to compute less useful. Nevertheless, it is interesting whether compression of a 7- or 8-second history should be applied when designing lexical, prosodic or other linguistic features, for prediction and for other tasks.

The segmentation for Switchboard-1 (but not for Spontal) was obtained using forced alignment [8]. This obviously corresponds to a better segmentation than would be available in a real prediction system. However, just as in language modeling in automatic speech recognition (ASR), the goal here was to capture the true distributions of n-grams; language models in ASR are not trained using hypotheses. On the other hand, predicitons based on a conditioning context built using detected rather than reference speech activity may behave differently, and may merit the design of a different windowing scheme which is more tolerant of frame errors.

The generalization of the modeling framework, inclusive of history compression, to other corpora, suggests that it may now be possible to conduct extensive comparisons of turn-taking across multiple languages, speaking styles, and group sizes in a quantifiable way. The modeling framework could also be applied in an analytical manner to detect whether — from a turn-taking point of view — spoken dialogue systems behave as humans do in the same settings.

In summary, we find that speech activity as far back as 7 seconds, of both the target speaker and their interlocutor, is relevant to the prediction of target speaker's incipient speech. This limit was discovered using a stochastic n-gram modeling framework by allocating a fixed number of model parameters to achieve an optimal trade-off between history precision and range. The resulting history compression profile is shown to be nearly logarithmic, and to generalize across conversational corpora in different languages and styles. The approach may be applicable to other modeling frameworks, may yield more accurate and faster detection and prediction systems, and should inform the design of other feature types which describe recent behavior.

5. REFERENCES

[1] J. Jaffe, L. Cassotta, and S. Feldstein, "Markovian model of time patterns of speech," *Science (New Series)*, vol. 144, no. 3620, pp. 884–886, 1964.

[2] J. Jaffe, S. Feldstein, and L. Cassotta, "Markovian models of dialogic time patterns," *Nature*, vol. 216, pp. 93–94, 1967.

[3] P. Brady, "A model for generating on-off speech patterns in two-way converation," *Bell Systems Technical Journal*, vol. 48, no. 9, pp. 2445–2472, 1969.

[4] T. Wilson, J. Wiemann, and D. Zimmerman, "Models of turn-taking in conversational interaction," *Journal of Languge and Social Psychology*, vol. 3, no. 3, pp. 159–183, 1984.

[5] A. Raux and M. Eskenazi, "Finite state turn taking model for spoken dialog systems," in *Proc HLT*, Boulder CO, USA, 2009, pp. 629–637.

[6] K. Laskowski, J. Edlund, and M. Heldner, "Incremental learning and forgetting in stochastic turn-taking models," in *Proc. INTERSPEECH*, Firenze, Italy, 2011, pp. 2069–2072.

[7] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*, San Francisco CA, USA, 1992, vol. 1, pp. 517–520.

[8] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of SWITCHBOARD," in *Proc. ICSLP*, Sydney, Australia, 1998.

[9] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture," in *Proc. LREC*, La Valletta, Malta, May 2010, ELRA, pp. 2992–2995.

[10] M. Heldner, J. Edlund, A. Hjalmarsson, and K. Laskowski, "Very short utterances and timing in turn-taking," in *Proc. INTERSPEECH*, Firenze, Italy, 2011, pp. 2837–2840.

[11] K. Laskowski, *Predicting, Detecting and Explaining the Occurrence of Vocal Activity in Multi-Party Conversation*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh PA, USA, 2011.