EXPLOITING LOUDNESS DYNAMICS IN STOCHASTIC MODELS OF TURN-TAKING

Kornel Laskowski

Carnegie Mellon University, Pittsburgh PA, USA
Voci Technologies, Inc., Pittsburgh PA, USA

ABSTRACT

Stochastic turn-taking models have traditionally been implemented as N-grams, which condition predictions on recent binary-valued speech/non-speech contours. The current work re-implements this function using feed-forward neural networks, capable of accepting binary- as well as continuous-valued features; performance is shown to asymptotically approach that of the N-gram baseline as model complexity increases. The conditioning context is then extended to leverage loudness contours. Experiments indicate that the additional sensitivity to loudness considerably decreases average cross entropy rates on unseen data, by 0.03 bits per framing interval of 100 ms. This reduction is shown to make loudness-sensitive conversants capable of better predictions, with attention memory requirements at least 5 times smaller and responsiveness latency at least 10 times shorter than the loudness-insensitive baseline.

Index Terms— Interaction models, neural networks, prosody, spoken dialogue systems.

1. INTRODUCTION

Among predictive models of turn-taking [1], those offering predictions at *every* instant are arguably the most relevant to the design of dialogue systems capable of truly human-like responsiveness. Although studied for many decades [2, 3, 4, 5, 6, 7, 8], these models continue to exhibit an important limitation: their implementation as N-grams circumscribes their direct applicability to only discrete-valued representations of conditioning context. This limitation has made it hard to study the impact of quantities which are continuous-valued (e.g., loudness or pitch), independently of higher-level linguistic landmarks or assumptions. Prediction using continuous-valued features has consequently been explored almost exclusively for detecting speaker change, at the ends of semantically complete utterances [9] or of contiguous intervals of speech [10], rather than continuously over all ongoing instants in a conversation.

The current work is concerned with eliminating this shortcoming, by re-implementing turn-taking models using feed-forward neural networks (NNs), a popular non-linear regression methodology. The presented experiments demonstrate that NNs achieve asymptotically identical performance in terms of average cross entropy on unseen data, given the same truncation of speech/non-speech history for prediction conditioning. More importantly, however, is the fact that NNs easily accept continuous-valued features. As a particular example of such a feature, the current work augments stochastic turn-taking modeling with sensitivity to *loudness*; its focus is to answer three questions:

- *How should models condition predictions of incipient speech on past loudness estimates in order to maximally reduce average cross entropy?*
- *What is the lowest average cross entropy achievable with loudness-augmented turn-taking models?* and
- *What is the likely impact of the observed average cross entropy reductions?*

Experiments show that contours of loudness, approximated by normalized per-frame log-energy, should be concatenated with speech activity trajectories in feature space rather than in model space (as in [6]), in order to give models the opportunity to leverage cross-stream correlations; it appears that the most relevant information is found in audio frames which are *both* speech *and* very quiet. The absolute reduction in average cross entropy obtained using this approach, on unseen data consisting of 200 telephone conversations, is 0.031 bits per 100 ms frame of audio, a large improvement when compared to past research [7, 8]. It is shown that the nominal impact of this improvement is to decrease requisite conversant attentiveness, by reducing memory requirements by a factor of at least 5, and to increase responsiveness, by enabling predictions of similar or better quality to be made at least 10 times sooner.

2. N-GRAMS OVER CHRONOGRAMS

A *chronogram* is a time-aligned per-frame speech activity sequence for all participants to a conversation. It is merely a convenient formalism, described in [2], and here denoted $\mathbf{Q}$. In the current two-party setting, $\mathbf{Q}$ is a $K \times T$ matrix with $K \equiv 2$, where T is the number of frames. As in [7, 8], the frame step is 100 ms, which corresponds approximately to half of a syllable at nominal speaking rates. An example of $\mathbf{Q}$, with each entry $\mathbf{q}_t\left[k\right] \in \{\square, \blacksquare\}$, with $\square$ and $\blacksquare$ representing non-speech and speech, respectively, and with $1 \leq t \leq T$ and $1 \leq k \leq K$, is

$$\mathbf{Q} \quad = \quad \begin{bmatrix} \cdots & \blacksquare\blacksquare\blacksquare\blacksquare\square\square\square\square\blacksquare & \cdots \\ & \square\square\square\blacksquare\blacksquare\blacksquare\blacksquare\square\square & \end{bmatrix} . \tag{1}$$

A *stochastic turn-taking model* $\mathbf{\Theta}$ is a device which yields, at any instant t, the probability that a target participant in $\mathbf{Q}$ will be in state $\blacksquare$. A target participant is either that represented by the top row ($k = 1$) in $\mathbf{Q}$, or that represented by the bottom row ($k = 2$). In the stochastic turn-taking modeling tradition, this probability of $\blacksquare$ can be conditioned by the target participant's own past speech activity states, or by both participants' past speech activity states. The latter case is a form of Markov chain coupling, as $\mathbf{\Theta}$ assumes the two participants to be *conditionally independent* (CI) at instant t (independent at t but conditioned on the joint past). In the former case, the two participants are assumed to be *unconditionally independent* (UI), and no coupling is possible.

To date, stochastic turn-taking models have been implemented as N-grams. This is a natural choice, since the conditioning context is composed of binary-valued events, whose composition is therefore also discrete-valued. Model performance on a dataset is expressed

978-1-4673-5126-3/12/$31.00 ©2012 IEEE

as average cross entropy, in bits, mirroring the use of perplexity in language modeling. The baseline N-gram models in this work were described in [7, 8]: they are Jelinek-Mercer-smoothed 11-grams in the UI case and 21-grams in the CI case, when conditioned on $\tau = 10$ most recent frames for the target participant and for both participants, respectively.

3. ALTERNATIVE BERNOULLI ESTIMATORS

The N-gram model described above is seen to implement a look-up table. To "compute" the probability $P(\mathbf{q}_t[k] = \blacksquare)$, the model looks at preceding instants $\tau < t$ in the $\mathbf{Q}$ under test, for either only the kth row or both rows, depending on whether $\mathbf{\Theta}$ is a UI or a CI model, respectively. This context, in its entirety, is a discrete-valued "key", using which the model retrieves a single stored value. The retrieved value is the Bernoulli probability that $\mathbf{q}_t[k] = \blacksquare$.

An alternative means of estimating these Bernoulli probabilities is to *actually* compute them. Namely, under a mapping $\square \mapsto 0$ and $\blacksquare \mapsto 1$, $\mathbf{\Theta}$ may be implemented as a regressor of arbitrary complexity whose output matches the N-gram estimate to desired precision. A first aim of the current work is to demonstrate the practical equivalence of N-gram and regressor implementations of $\mathbf{\Theta}$. However, the main motivation for exploring alternative estimators is to allow $P(\mathbf{q}_t[k] = \blacksquare)$ to be conditioned on continuous-valued, rather than only discrete-valued, "keys".

The most popular non-linear regression scheme in use today is logistic regression [11], which has been argued to be particularly well-suited to the "resolution" of overlap in multi-party conversation [12]. A more flexible scheme, employed in the current work, is offered by feed-forward neural networks (NNs), whose output layer activation function is the sigmoid function used in logistic regression. The appropriate objective function to minimize during learning in this case is the cross entropy error [13]. All NNs presented consist of a single hidden layer, with a variable number J of hidden units. Learning is accelerated via a second-order technique known as scaled conjugate gradient search [14].

4. APPROXIMATING LOUDNESS WITH ENERGY

Research in psychoacoustics has produced many methods for approximating perceptual loudness [15] using spectral decomposition, spectral masking, and/or temporal masking. These models appear more applicable to single-tone and/or stationary signals than to human speech. In the current work, loudness is instead approximated by signal energy, as has become standard in speech technology. The correlation between loudness and signal energy is sufficiently strong that the two terms are often used interchangeably in the literature, despite the fact that authors use different definitions of signal energy.

In the current work, for every conversation, and each channel k separately, the energy e_t in the tth frame is computed using

$$e_t[k] \quad = \quad \log_{10} \sum_{u=-\lfloor W/2 \rfloor}^{+\lceil W/2 \rceil - 1} w[u] \cdot x^2[Nt+u] \ , \qquad (2)$$

where $x[\cdot]$ is the discrete-time pressure signal (sampled at 8 kHz). W is the frame size (of 1600 points, corresponding to 200 ms), N is the frame step (of 800 points, corresponding to 100 ms), and $w[\cdot]$ is a 200-ms Hann window normalized to unity area and centered on zero. e_t is seen to be the logarithm of a weighted sum squared amplitude[1]. This arrangement also ensures that the dimensions of $\mathbf{E} \equiv \{e_t[k]\}$, for each conversation, are identical to those of that conversation's $\mathbf{Q}$, namely $K \times T$.

5. EXPERIMENTS

5.1. Data

Experiments are conducted using the Switchboard-1 Corpus, as re-released in 1997 [16]. It consists of 2435 telephone conversations, each approximately 10 minutes in duration. The corpus was divided into three speaker-disjoint sets in [8], such that TRAINSET, DEVSET, and TESTSET consist of 762, 227, and 199 conversations, respectively. During that process, it was not possible to allocate 1247 conversations because their two speakers had already been placed in different sets. Reference speech/non-speech segmentations were used, as elsewhere in speech technology when training prior probability models (e.g. language models). The available forced alignments [17] for both conversation sides were used to construct $\mathbf{Q}$.

5.2. Speech Activity Features

The first experiments compare the performance of the baseline smoothed N-gram models of turn-taking to feed-forward neural networks, with identical context. The latter are trained "to completion", using as many iterations of scaled conjugate gradient search as it takes for the TRAINSET error difference to reach zero[2]. The experiment is repeated using $\tau \in \{1, 2, \ldots, 10\}$ most-recent frames of history, with $J \in \{1, 2, 4, 8\}$ units in the NN hidden layer. The DEVSET results are shown in Figure 1 for both the UI and the CI models; the DEVSET was used to select optimal smoothing parameters for the N-gram models, but was not used for anything other than model scoring in the NN cases.

It is seen that NNs with a single hidden layer unit are outperformed by the corresponding UI and CI baseline models. However, as the number of hidden units grows, NN performance asympotically approaches that of the baselines. By $J = 8$, the UI NN achieves near-identical performance to the UI N-gram (the difference is < 0.00005 bits); in the CI case, the difference is negligible (< 0.0002 bits), and likely to narrow further for still-larger values of J.

It is worth mentioning that whereas the N-gram models at $\tau = 10$ must explicitly store all probabilities, and therefore consist of $2^{10} = 1024$ and $2^{20} = 1048576$ free parameters for the UI and CI cases, respectively, the comparable NN models are much smaller. For the UI NN with $\tau = 10$ and $J = 8$, the number of free parameters is $J \cdot (10 + 1) + 1 \cdot (J + 1) = 97$; for the CI NN, it is $J \cdot (20 + 1) + 1 \cdot (J + 1) = 177$. This parsimony comes of course at the cost of the much higher time complexity of iterative parameter estimation.

5.3. Loudness Features

The next experiments contrast the performance of speech activity features with that of loudness features, using the NN approach of the preceding subsection. Energies are computed using Equation 2 independently for both channels in each conversation, and the resulting $\mathbf{E}$ is used in place of $\mathbf{Q}$. For expediency, NNs are not trained

[1] Popular variations include window shape, the ratio W/N, window normalization to unity sum squared amplitude, and choice of multiplicative factor; for example, a factor of 10 would nominally result in the popular dB scale of sound pressure level. Most of these variations are affine transformations of Equation 2, and therefore lead to equivalent NN models whose hidden-layer activation function is a biased dot product.

[2] Effectively, to underflow in IEEE 64-bit floating-point precision.

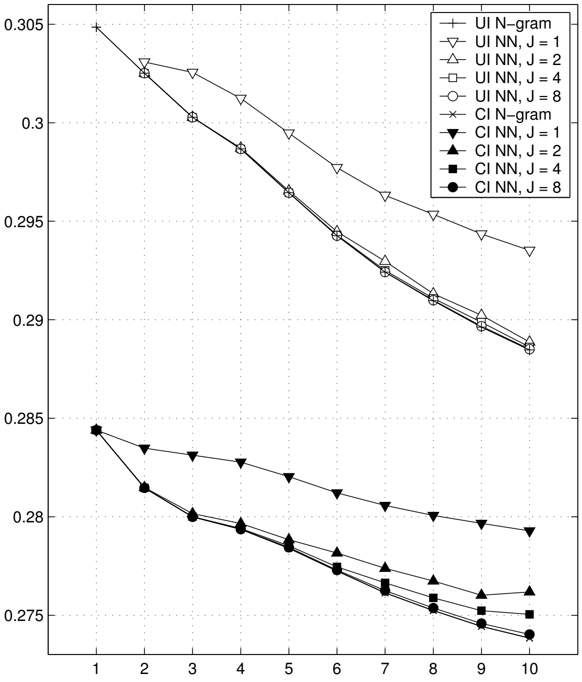


Fig. 1. DEVSET cross entropy (along the y-axis, in bits) as a function of the size of the conditioning context in 100-ms frames (along the x-axis), for the N-gram and for several neural network (NN) models, using only binary reference speech activity. J represents the number of NN hidden units.

to completion; instead, 100 iterations of search are executed for increasingly larger subsets of TRAINSET. The 1st subset consists of every 1024th exemplar, the 2nd of every 512th exemplar, and so on until the 11th subset, which consists of all TRAINSET exemplars.

All NNs are exposed to a conditioning history of $\tau = 10$ 100-ms frames, with $J \in \{1, 2, 4, 8, 16, 32, 64\}$. The optimal value of J is selected by minimizing cross entropy on DEVSET. The achieved UI and CI cross entropies, of 0.481 and 0.381, respectively, are shown in the second row of Figure 2 as "E". The performance of $\mathbf{Q}$ features, using the same training regimen, is shown in the first row as "Q". As can be seen, raw energy features yield substantially worse performance that do speech activity features. Evidently, although $\mathbf{E}$ contains dynamic loudness information, the NNs are unable to infer which context frames are speech and which are non-speech, based on the context energies alone.

To assess whether this is due to channel differences among the conversations, the above experiment is repeated after computing several standard per-channel normalizations of e_t. Among these are: $e'_t \equiv e_t - \mu$ (mean subtraction), where μ is the global channel mean; $e'_t \equiv e_t - \min_\tau e_\tau$ (minimum subtraction), where $\min_\tau e_\tau$ is the global channel minimum; $e'_t \equiv e_t - \max_\tau e_\tau$ (maximum subtraction), where $\max_\tau e_\tau$ is the global channel maximum; and $e'_t \equiv e_t/\sigma - \mu/\sigma$ (Z-normalization), where μ and σ are the global channel mean and global standard deviation, respectively[3].

The felicity of normalization, relative to the performance of the raw, unnormalized e_t, is shown in the third through sixth rows of Figure 2. Minimum subtraction appears to unacceptably degrade performance, most likely due to the effect of zeroing out intervals of conversation sides upon a participant's request. For UI models, all three of the remaining methods improve performance, with mean subtraction offering the lowest cross entropies. However, for CI models, per-channel normalization hurts. This may be due to channel coupling [18], potentially calling for normalization schemes which treat the two channels jointly.

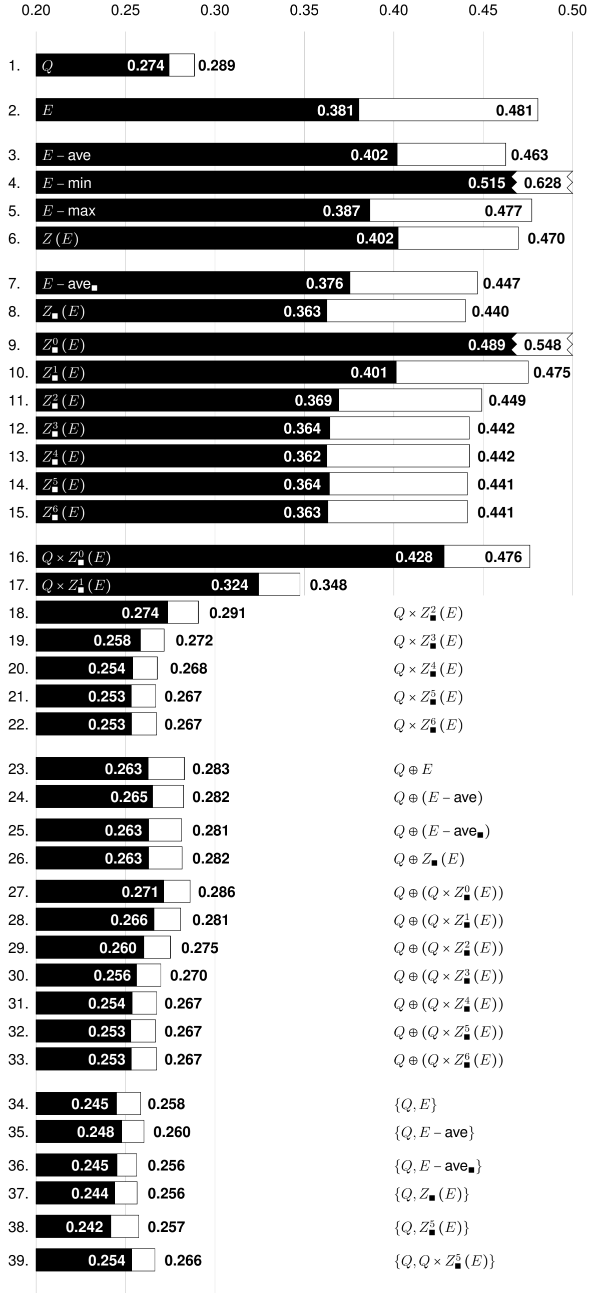


Fig. 2. DEVSET cross entropies (along the x-axis, in bits) for selected feature types. UI and CI models are shown in white and black, respectively.

[3] These normalizations are acausal, since the minimum, maximum, mean, and standard deviations are computed using all of the channel audio, including instants beyond the current instant $t < T$. The estimation of these quantities, in a truly on-line setting, is deferred to future work.

5.4. Normalizing Loudness Features Using Speech Activity

To eliminate the dependency on arbitrary signal energy floors, mean subtraction and Z-normalization are carried out using means and standard deviations computed over only speech frames. This conditions the normalized $\mathbf{E}$ on $\mathbf{Q}$. The DevSet results are shown in the 7th and 8th rows of Figure 2, as $E - \text{ave}_\blacksquare$ and $Z_\blacksquare(E)$, respectively. They demonstrate that statistics computed while ignoring non-speech frames are more robust than global statistics, as both normalization types lead to both UI and CI improvement relative to unnormalized energy, as well as relative to rows 3 and 6.

Finally in this subsection, $Z_\blacksquare$-normalization is modified to help the NNs infer $\mathbf{Q}$ from $\mathbf{E}$. Namely, energies below a predefined number of standard deviations $\sigma_\blacksquare$ are zeroed, while simultaneously ensuring that all other energies are positive:

$$e'_t \equiv \begin{cases} \alpha + \frac{e_t}{\sigma_\blacksquare} - \frac{\mu_\blacksquare}{\sigma_\blacksquare} & \text{if } e_t \geq \mu_\blacksquare + \alpha \cdot \sigma_\blacksquare \\ 0 & \text{otherwise} \end{cases} . \tag{3}$$

The results of this manipulation, referred to here as $Z_\blacksquare^\alpha$-normalization, are shown in rows 9 through 15 of Figure 2, corresponding to $\alpha \in \{0, 1, 2, 3, 4, 5, 6\}$.

It appears that $\mathbf{E}$ is not sufficiently well correlated with $\mathbf{Q}$, on average, to allow NNs to recover the performance observed in the first row of Figure 2. As fewer and fewer frame energies are zeroed, performance asympotically approaches that of the $Z_\blacksquare(E)$ normalization in row 8, for both UI and CI models. The utility of loudness (present in $\mathbf{E}$) appears to be smaller than that of speech/non-speech (present in $\mathbf{Q}$ but evidently not inferrable from $\mathbf{E}$ alone).

5.5. Combining Loudness and Speech Activity Features

Loudness and speech/non-speech information can be more directly combined by modifying Equation 3 to explicitly contain $\mathbf{Q}$:

$$e'_t \equiv \begin{cases} \delta(q_t, \blacksquare) \cdot \left(\alpha + \frac{e_t}{\sigma_\blacksquare} - \frac{\mu_\blacksquare}{\sigma_\blacksquare} \right) & \text{if } e_t \geq \mu_\blacksquare + \alpha \cdot \sigma_\blacksquare \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\delta(\cdot)$ is the Kronecker delta. The performance of the resulting $\mathbf{E}'$ is shown as $Q \times Z_\blacksquare^\alpha(E)$ in rows 16 through 22 in Figure 2. Rows 16 through 22 should be compared to rows 9 through 15, which differ only in multiplication by $\delta(q_t, \blacksquare)$.

Equation 4 is seen to yield dramatically improved results. At $\alpha = 2$ (row 18), when all speech frames whose e_t is $2\sigma_\blacksquare$ below $\mu_\blacksquare$ and all non-speech frames are zeroed, performance is nearly identical to that in the first row, for $\mathbf{Q}$ alone. Increasing α and thereby retaining more of the frames whose e_t is below $\mu_\blacksquare$ asymptotically leads to an improvement of approximately 0.021 bits (bigger than the 0.015-bit gap between UI and CI models for $\mathbf{Q}$ alone, in row 1). It appears that exploitable loudness information is found in speech frames which are quieter than 99% of speech frames, namely those most easily confounded with non-speech frames.

5.6. Combining Loudness and Speech Activity Models

Second, $\mathbf{Q}$ and $\mathbf{E}$ information is combined using model fusion, by linearly interpolating the probabilities returned by the $\mathbf{Q}$ models and the normalized- or unnormalized-$\mathbf{E}$ models. DevSet cross entropies are minimized to select the optimal interpolation weights in each pair-wise fusion case. The results are indicated by a $\oplus$ symbol in rows 23 through 33 in Figure 2.

Interpolation with Q appears to improve the performance of E (row 2 versus row 23), $E - \text{ave}$ (row 3 versus row 24), $E - \text{ave}_\blacksquare$ (row 7 versus row 25), and $Z_\blacksquare(E)$ (row 8 versus 26), yielding cross entropies which are below that achieved with $\mathbf{Q}$ alone (in row 1) by approximately 0.01 bits. When interpolating Q with $Z_\blacksquare^\alpha$-normalization, performance approaches that achieved for $\alpha = 6$ (in row 33), but does not exceed that achieved by $Z_\blacksquare^\alpha$-normalization alone.

5.7. Concatenating Loudness and Speech Activity Features

Finally, feature fusion as opposed to model fusion is attempted, by concatenating $\mathbf{Q}$ and $\mathbf{E}$ and training new (and twice as large) models. Because of the higher computational complexity, feature fusion with $\mathbf{Q}$ was performed for only a select number of $\mathbf{E}$-derived feature sets, shown in rows 34 through 39 of Figure 2.

Feature fusion is seen to yield the best results of all. For UI models, fusion of $\mathbf{Q}$ with $E - \text{ave}_\blacksquare$ yields an improvement over the first row in the figure of 0.033 bits; for CI models, the best combination is of $\mathbf{Q}$ with $Z_\blacksquare^\alpha$-normalized $\mathbf{E}$, using $\alpha = 5$, and resulting in an improvement over the first row of 0.032 bits. Interestingly, combining $\mathbf{Q}$ with $Z_\blacksquare^5(E)$ is better than combining $\mathbf{Q}$ with $Q \times Z_\blacksquare^5(E)$. This suggests that the NNs are exploiting loudness dynamics during non-speech frames. These could be informative of inspiration or expiration systematically surrounding speech [19], of other systematic non-verbal vocalizations from each channel's speaker [20, 21], or of crosstalk from the interlocutor channel [18].

6. DISCUSSION

6.1. Generalization

Of importance to the establishment of any claim is whether the performance observed for DevSet generalizes to unseen data. DevSet was used to select the optimal number of hidden units as well as the interpolation parameters in Subsection 5.6, and the observed gains may be optimistic indicators of performance.

Figure 3 indicates that this is not so. Virtually all qualitative observations made in the preceding section apply equally well to the completely held-out TestSet. Overall, TestSet appears to be slightly easier. Feature fusion via concatenation yields the best results: for UI models, the best performance in row 36 is 0.033 bits below that in the first row. For CI models, the best performance is found in row 38, and is lower than that in the first row by 0.034 bits.

6.2. What Loudness-Sensitive Models Learn

NN models are easy to visualize via Hinton diagrams [22]. Figure 4 compares the first-layer weights in two sample UI NNs (with and without loudness) with $J = 4$ hidden units; first-layer biases and second-layer weights and biases are not shown. First-layer basis functions are shown from top to bottom in order of importance.

The figure shows that the better performing model, in (b), observes a sparser first-layer weight matrix, and that large-magnitude weights are concentrated on the most recent events. Despite ignoring very many of the instants that the NN in (a) is sensitive to, (b) is better performing. This suggests that sensitivity to loudness allows turn-takers to be less attentive: they can afford to make decisions without paying attention to and retaining longer historical contexts.

6.3. An Alternative Model Assessment Methodology

Throughout this paper, models have been quantized in terms of cross entropy in bits per 100 ms, and in terms of the number of bits fewer than is required by the baseline models which use only $\mathbf{Q}$. These measures are difficult to interpret by themselves, but they do allow

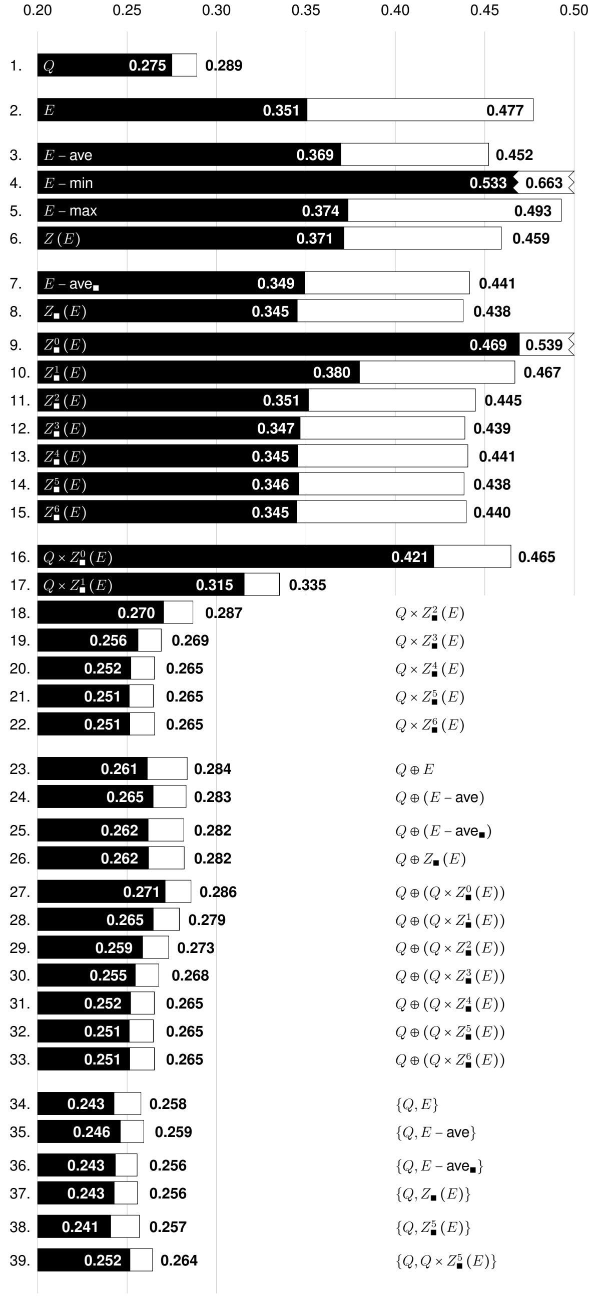
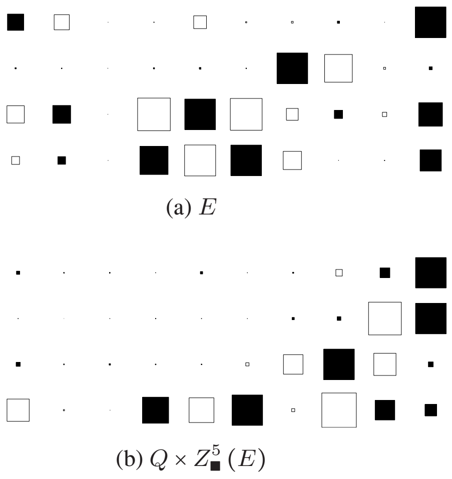


Fig. 3. TESTSET cross entropies (along the x-axis, in bits) for selected feature types. UI and CI models are shown in white and black, respectively.

for strict comparison *between* models. In particular, it is possible to ablate the capability of a better-performing model until it achieves performance quantitatively comparable to a baseline model. If the capability in question is the amount of historical context, a difference in cross entropies can be converted to the savings in requisite attentiveness and response latency.

Figure 5 shows just such an experiment, in which the best performing feature types for UI and CI models (feature-level fusion with Q of $E-$ave$_{\blacksquare}$ and $Z^{5}_{\blacksquare}(E)$, respectively) are exposed to shorter truncations of the conditioning history, e.g. $\tau \in \{1,2,3,4,5,6,7,8,9\} < 10$. The results are shown for DEVSET, but the trends for TESTSET are the same; all NNs employ $J = 64$ hidden units. Also shown are the baseline UI and CI model performance, with $J = 64$ and $\tau = 10$, with speech/non-speech features alone.



(a) E

(b) $Q\times Z^{5}_{\blacksquare}(E)$

Fig. 4. Hinton diagrams of the first-layer weights in two UI NN models with $\tau = 10$ 100-ms frames of context and $J = 4$ hidden units. Black indicates positive weights, white negative weights; dots correspond to near-zero values. Basis functions shown from top to bottom in order of importance (as indicated by the magnitude of the corresponding second layer weights); biases not shown. Rows are normalized such that the largest-valued weight in each row has a fixed size. Age of context shown from left to right: the rightmost column corresponds to the most recent frame (100 ms ago), the leftmost column to the least recent (1 s ago).

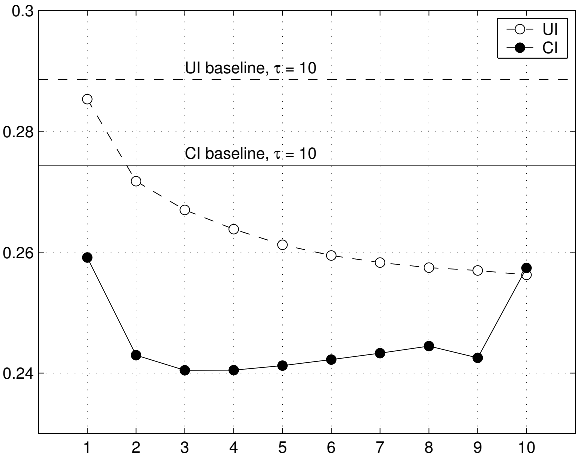


Fig. 5. DEVSET cross entropy (along the y-axis, in bits) as a function of the size of the conditioning context in 100-ms frames (along the x-axis), for the best-performing combined speech activity and loudness models. The baselines shown are for a conditioning context of 10 frames, achieved with a NN model using speech activity only.

What is apparent is that even a single frame ($\tau = 1$) of both speech/non-speech and normalized energy yield lower cross entropies than do $\tau = 10$ frames of speech/non-speech context alone, in both the UI and CI cases. This means that conversational agents needing to predict incipient speech will be less surprised with 5 times less storage of historical data. Similarly, by paying attention to loudness, conversational agents urgently needing to say something need only wait for 100 ms of context rather than for 1 s, in order to do as well if not better. Of course, they can do far better still by attending to more than just the most recent frame, as can be observed in Figure 5[4].

7. CONCLUSIONS

This work has demonstrated that traditional stochastic turn-taking models, namely N-grams over past speech activity, can be successfully implemented using neural networks (NNs). NN models were shown to yield asympototically identical performance, as the number of hidden units is increased, with considerably fewer parameters.

More importantly, NNs open up the possibility of conditioning predictions on continuous-valued quantities, such as loudness, pitch, and spectral content. Experiments showed that augmentation with sensitivity to loudness contours, approximated by standard log-energy estimates, reduces average cross entropy on unseen data by at least 0.031 bits per 100 ms frame. It was argued that this reduction entails savings in requisite attentiveness for conversants, by reducing memory capacity requirements by a factor of at least 5. Similarly, sensitivity to loudness enables participants to be at least 10 times more responsive, needing to wait only a tenth of the time to produce better predictions than when loudness is ignored.

8. ACKNOWLEDGMENTS

The author is grateful to Mattias Heldner and Jens Edlund for suggestions and helpful discussion. The work was funded in part by the Riksbankens Jubileumsfond (RJ) project *Samtalets Prosodi*. Computing resources at Carnegie Mellon University were made accessible courtesy of Qin Jin.

9. REFERENCES

[1] 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, doi:10.1177/0261927X8400300301.

[2] E. Chapple, "The interaction chronograph: Its evolution and present application," *Personnel*, vol. 25, pp. 295–307, 1949.

[3] 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.

[4] J. Jaffe, S. Feldstein, and L. Cassotta, "Markovian models of dialogic time patterns," *Nature*, vol. 216, pp. 93–94, 1967, doi:10.1038/216093a0.

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

[6] A. Raux, *Flexible Turn-Taking for Spoken Dialog Systems*, Ph.D. thesis, Carnegie Mellon University, 2009.

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

[8] K. Laskowski and E. Shriberg, "Corpus-independent history compression for stochastic turn-taking models," in *Proc. ICASSP*, Kyoto, Japan, March 2012, pp. 4937–4940, doi:10.1109/ICASSP.2012.6289027.

[9] L. Ferrer, E. Shriberg, and A. Stolcke, "Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog," in *Proc. ICSLP*, Denver CO, USA, 2002, vol. 3, pp. 2061–2064.

[10] K. Laskowski, J. Edlund, and M. Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems," in *Proc. ICASSP*, Las Vegas NV, USA, April 2008, pp. 5041–5044, doi:10.1109/ICASSP.2008.4518791.

[11] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics. Springer, New York NY, USA, 2004.

[12] K. Laskowski, M. Heldner, and J. Edlund, "Preliminaries to an account of multi-party conversational turn-taking as an antiferromagnetic spin glass," in *Proc. Workshop on Modeling Human Communication Dynamics at NIPS*, Whistler BC, Canada, December 2010, pp. 46–49.

[13] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York NY, USA, 1995.

[14] M. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993, doi:10.1016/S0893-6080(05)80056-5.

[15] H. Fastl and E. Zwicker, *Psychoacoustics*, Springer, 3rd edition, 2006.

[16] J. J. Godfrey and E. C. Holliman, *Switchboard-1 Release 2*, Number LDC97S62. Linguistic Data Consortium, Philadelphia PA, USA, 1997.

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

[18] D. Liu and F. Kubala, "A cross-channel modeling approach for automatic segmentation of conversational telephone speech," in *Proc. ASRU*, St. Thomas, US Virgin Islands, November 2003, pp. 333–338, doi:10.1109/ASRU.2003.1318463.

[19] D. McFarland, "Respiratory markers of conversational interaction," *Journal of Speech, Language and Hearing Research*, vol. 44, pp. 128–143, February 2001, doi:10.1044/1092-4388(2001/012).

[20] S. Duncan and G. Niederehe, "On signalling that it's your turn to speak," *Journal of Experimental Social Psychology*, vol. 10, no. 3, pp. 234–247, May 1974, doi:10.1016/0022-1031(74)90070-5.

[21] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," *Speech Communication*, vol. 53, no. 1, pp. 23–35, January 2011, doi:10.1016/j.specom.2010.08.003.

[22] G. Hinton, J. McClelland, and D. Rumerhart, *Parallel Distributed Representations: Explorations in the Microstructure of Cognition*, chapter 3: Distributed representations, pp. 77–109, MIT Press, Cambridge MA, USA, 1986.

[4] The upswing in the CI curve for $\tau > 3$ is also present in the TRAINSET curve, suggesting that the training regimen of 11×100 iterations of scaled conjugate gradient search may be insufficient for large values of J, yielding considerably undertrained networks. Fully trained NNs are expected to do better on all three datasets.