A Scalable Method for Quantifying the Role of Pitch in Conversational Turn-Taking

Kornel Laskowski[1,2] and **Marcin Włodarczak**[1] and **Mattias Heldner**[1]
[1] Stockholm University, Stockholm, Sweden
[2] Voci Technologies, Inc., Pittsburgh PA, USA

Abstract

Pitch has long been held as an important signalling channel when planning and deploying speech in conversation, and myriad studies have been undertaken to determine the extent to which it actually plays this role. Unfortunately, these studies have required considerable human investment in data preparation and analysis, and have therefore often been limited to a handful of specific conversational contexts. The current article proposes a framework which addresses these limitations, by enabling a scalable, quantitative characterization of the role of pitch throughout an entire conversation, requiring only the raw signal and speech activity references. The framework is evaluated on the Switchboard dialogue corpus. Experiments indicate that pitch trajectories of both parties are predictive of their incipient speech activity; that pitch should be expressed on a logarithmic scale and Z-normalized, as well as accompanied by a binary voicing variable; and that only the most recent 400 ms of the pitch trajectory are useful in incipient speech activity prediction.

1 Introduction

Intonation is generally held to play an integral role in the phonetic realization of turns and in the prediction of more talk (see e.g. (Bögels and Torreira, 2015) for a review). There is broad consensus that flat pitch segments are associated with turn-holding and that rising or falling pitch segments are associated with turn-yielding (Bögels and Torreira, 2015; Caspers, 2003; Duncan, 1972; Edlund and Heldner, 2005; Ford and Thompson, 1996; Heldner et al., 2009; Heldner and Włodarczak, 2015; Hjalmarsson, 2011; Jefferson, 1984; Kane et al., 2014; Koiso et al., 1998; Laskowski et al., 2009; Local et al., 1986; Selting, 1996; Yanushevskaya et al., 2014; Zellers, 2013, 2017). Studies considering finer-grained categories of the pitch contour (Gravano and Hirschberg, 2011; Wennerstrom and Siegel, 2003), additionally including slowly rising and slowly falling pitch, have tended to corroborate those findings. Furthermore, they indicate that the endpoint of a pitch segment is relevant, associating segments reaching the top or bottom of a speaker's range with turn-yielding and those ending near the middle of the range with turn-holding.

The converging results of so many studies are astonishing given the methodological differences between them, with regard to the speech material (spontaneous vs. task-oriented) and to the pitch-contour categorization method (perceptual judgements, acoustic measurements, or phonologically motivated categories). Perhaps more importantly, the studies in question differ in how the pitch contour is parametrized (e.g. perceptual stylization, functional data analysis, linear or polynomial curve fitting, linear or logarithmic scale), how far back in the speech interval relevant pitch cues are to be found, as well as how cues are evaluated (e.g. perceptually vs. statistically).

It is therefore not very surprising that work which has tried to verify the above claims with acoustic measurements of fundamental frequency (F0) has also produced some mixed results (see e.g. (Zellers, 2017; Walker, 2017) for reviews). A variety of explanations for these mixed results are believed to exist. First, it has been hypothesized that non-pitch cues may play a more important role than do pitch cues (e.g. (Local and Walker, 2012; Walker, 2017; Zellers, 2017). Second, it is possible that the role of intonation varies with the communicative situation, and that it is strongly dependent on the number of participants, whether the participants have eye contact, whether the participants know one another, etc. Finally, there may be considerable language-, dialect-, and domain-specific differences in the role of pitch in turn-

Proceedings of the SIGDial 2019 Conference, pages 284–292
Stockholm, Sweden, 11-13 September 2019. ©2019 Association for Computational Linguistics

taking. At the present time, these explanations continue to be mere hypotheses which — owing to the many methodological differences in published work — cannot be easily evaluated.

The main focus of the current article is to render the evaluation and comparison of such hypotheses tractable, if not outright easy. A key requirement is that the proposed method be *scalable*, i.e. capable of ingesting sufficiently large quantities of conversational material to generate representative results. This in turn requires that it not rely on time-consuming, often-contentious annotation of either turn or pitch phenomena — authors of existing research do not always agree on what constitutes a turn, for example. Furthermore, the method needs to be *quantitative* if it is to permit strict comparison. The method proposed in the current article is both scalable and quantitative; it relies only on the availability of the raw signal and accurate speech activity references, per conversation and per conversation-side. It is presented in Section 3.

To evaluate the method itself, the current article asks three key questions of a large, oft-studied corpus of telephone conversations in English (described in Section 2). These questions are:

Q1. *Can attention to pitch reduce the average surprise of incipient speech activity?*
Q2. *What is the optimal representation of pitch for a speech prediction system?*
Q3. *How far back into the past should a pitch-sensitive method look?*

Experiments described in Section 4 demonstrate that *Q1* can be answered in the affirmative, that binary voicing and Z-normalized log-pitch offer the best results when used together (*Q2*), and that only the most recent 400 ms of pitch history are sufficient (*Q3*). Furthermore, the proposed system is able to answer *Q1* and *Q3* in a nearly fully-automated fashion, for evidently any corpus; the answer to *Q2* may require human-mediated investigations, for which the proposed system provides a suitable and convenient framework.

2 Data

Experiments used the Switchboard-1 Corpus, as re-released in 1997 (Godfrey and Hollimann, 1997). The corpus consists of 2435 dyadic telephone conversations, each approximately 10 minutes in duration. It was iteratively divided into three speaker-disjoint sets as in (Laskowski and Shriberg, 2012), such that TRAINSET, DEVSET, and TESTSET consist of 762, 227, and 199 conversations, respectively. During the division process, it was not possible to allocate 1247 of the Switchboard-1 conversations, because each of their two speakers had already been placed in different sets. Forced alignments of the manually transcribed words (used as discussed in Subsection 3.2) for both sides of the conversation were provided in (Deshmukh et al., 1998).

3 Methods

This article proposes a means of quantifying the extent to which pitch, represented in a variety of ways, reduces the surprise induced by observing the temporal distribution of speech in unseen conversations. Such a means involves a probabilistic formulation of the problem (Subsection 3.1), a method for obtaining instantaneous binary speech activity (Subsection 3.2), a method for measuring pitch (Subsection 3.3), and a model for approximating the probabilities given those features, together with a metric for quantifying model performance (Subsection 3.4).

3.1 Stochastic Turn Taking

As in (Laskowski, 2012), the methodology employed here relies of forming a probability distribution over the side-attributed speech activity in entire dyadic conversations. This eliminates a dependency on the specific definition of a turn; the resulting probability models attempt to account for all speech, effectively marginalizing out alternative definitions of what turns are and where they start and end.

The most direct means of modeling conversations for this purpose is to discretize their temporal extent; here, a frame- step and size of 100 ms is used, representing approximately half of a normative syllable. Such discretization results in a $K \times N$ *chronogram* for each conversation, ie.

$$\mathbf{Q} = \begin{bmatrix} \cdots & \blacksquare\blacksquare\blacksquare\blacksquare\square\square\square\square\blacksquare & \cdots \\ & \square\square\square\blacksquare\blacksquare\blacksquare\blacksquare\square\square & \end{bmatrix}, \quad (1)$$

where the kth row, $1 \leq k \leq K$, represents the speech activity of one of the $K = 2$ sides to the conversation, and each column $\mathbf{q}_n$, $1 \leq n \leq N$, represents one 100-ms interval. Each $\mathbf{q}_n[k] \in \{\square, \blacksquare\} \equiv \{0, 1\}$, indicating that the kth party is either not-speaking or speaking in frame n, respectively.

The probability $\mathcal{P}$ of a given $\mathbf{Q}$ is then given by

$$\mathcal{P}(\mathbf{Q}) = \prod_{n=1}^{N} \mathcal{P}\left(\mathbf{q}_n \middle| \mathbf{q}_1^{n-1}\right) \quad (2)$$

$$\approx \prod_{n=1}^{N} \mathcal{P}\left(\mathbf{q}_n \middle| \mathbf{q}_{n-\tau}^{n-1}\right) \quad (3)$$

$$\approx \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{P}\left(\mathbf{q}_n[k] \,\middle|\, \mathbf{q}_{n-\tau}^{n-1}\right) \quad , \quad (4)$$

where Equation 3 represents a Markovian truncation of the history to the most recent τ frames, and Equation 4 assumes that participants are conditionally independent of one another in any given frame, but dependent on their joint past $\mathbf{q}_{n-\tau}^{n-1}$. The term *target participant* is used to refer to that side of the conversation for which the interior factor on the right-hand-side of the equation is being evaluated; when evaluating the left-hand-side over all cells in $\mathbf{Q}$, each of the $K = 2$ participants becomes the target participant half the time.

In this framework, quantifying the impact of pitch — or any other side information available in $K \times N$ matrix form as $\mathbf{X}$ — entails comparing the probability in Equation 4 to

$$\mathcal{P}(\mathbf{Q}|\mathbf{X}) \approx \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{P}\left(\mathbf{q}_n[k] \,\middle|\, \mathbf{q}_{n-\tau}^{n-1}, \mathbf{x}_{n-\tau}^{n-1}\right) \quad .$$

By excluding the current and future $\mathbf{x}_n^N$ from the conditioning context, the factor $\mathcal{P}\left(\mathbf{q}_n[k] \,\middle|\, \mathbf{q}_{n-\tau}^{n-1}, \mathbf{x}_{n-\tau}^{n-1}\right)$ is observed to be a causal prediction.

3.2 Speech Activity

The above equation forms a probability density over speech activity $\mathbf{Q}$ that *actually happened*, rather than speech activity that can be measured. The most accurate means currently available for producing $\mathbf{Q}$ is to perform forced time-alignment of the kth participant's audio channel to the words spoken by that participant. The resulting word boundaries are then aligned to the 100-ms frame boundaries which define $\mathbf{Q}$, and each $\mathbf{q}_n[k]$, $1 \leq n \leq N$ and $1 \leq k \leq K$, is assigned to 1 if the kth participant was speaking for 50% or more of the temporal support of the nth frame.

3.3 Pitch

Pitch was extracted using the `get_f0` implementation available in the Snack Sound Toolkit (Sjölander, 2001). In order to avoid contagion from the future (pitch tracking uses context to smooth candidate per-frame fundamental frequency estimates), a separate pitch track was extracted for the τ-duration conditioning context of *every* frame n in every channel k of every conversation[1]. Snack's default frame step is 10 ms; the resulting sequence of 10-ms pitch estimates was then aligned to the 100-ms frames in $\mathbf{Q}$, yielding side-information $\mathbf{P}$. Each cell $\mathbf{p}_n[k]$ of $\mathbf{P}$ was assigned to the mean of those voiced 10-ms pitch estimates of the kth participant's speech which fell entirely within the temporal support of frame n; it therefore sufficed for only one 10-ms-frame to be deemed as voiced by Snack in order for the 100-ms frame in $\mathbf{P}$ to be considered voiced[2]; unvoiced frames in $\mathbf{P}$ were assigned to `NaN`.

Note that pitch computed as described above may exhibit doubling and halving errors; the exploration of the impact of (manually) corrected pitch is beyond the scope of the current article. Similarly, phenomena such as diplophonia and creakiness are not explicitly treated.

3.4 Models and Metrics

The prediction probabilities described in Subsection 3.1 were approximated using a feed-forward neural network

$$\mathcal{P}\left(\mathbf{q}_n[k] \,\middle|\, \mathbf{q}_{n-\tau}^{n-1}, \mathbf{x}_{n-\tau}^{n-1}\right) \approx f\left(\mathbf{q}_{n-\tau}^{n-1}, \mathbf{x}_{n-\tau}^{n-1}\right)$$

with one hidden layer of H tanh units[3], and one sigmoid output unit — representing the probability that the kth participant is speaking at frame n. For most experiments in the current article, $H = 8$. Note that the network has no recurrence since determining the exact extent of the usefully conditioning history is of primary interest. Network weights were trained on TRAINSET, using

[1] This brute-force and seemingly inefficient approach proved to have considerable impact on the numerical results presented in Section 4, indicating that basing incremental predictions on non-incremental pitch extraction would have been a form of cheating.

[2] Other policies were explored, notably that in which at least half of the 10-ms frames need to be voiced; the results exhibited the same trends as those reported here, although numerically the cross entropy rates were slightly larger. It appears that better predictions are possible when more of the 100-ms frames in $\mathbf{P}$ are deemed voiced, even when some of those cells are more sensitive to outliers in the underlying 10-ms pitch trajectory.

[3] Note that tanh activation units in the network implicitly map `NaN` features to zero. This approach is likely sub-optimal, but provides a well-understood and simple-to-train baseline for improvements like those described in (Laskowski, 2015).

1000 iterations of scaled conjugate gradient (SCG; (Møller, 1993)) descent — a second-order, deterministic rather than stochastic procedure.[4]

The appropriate objective function given a single sigmoid output unit is the cross entropy error (Bishop, 1995); it was used during SCG training as well as in the subsequent evaluation of trained models. Since, for any given conversation and participant, the evaluation of the model for a sequence of frames can be thought of as a causal prediction, during testing the error is henceforth referred to as the *cross-entropy rate*, and is expressed in bits per 100-ms frame.

4 Results

4.1 Representation

The first suite of experiments attempts to identify an optimal representation of pitch for the analysis task at hand. To put the ensuing results into perspective, the baseline is a system which excludes all pitch information; Figure 1 depicts as "Q^{τ}" the achieved cross entropy rate as a function of the number τ of past speech activity frames which comprise the conditioning context. As can be seen, the cross entropy rate exhibits a nearly linear decline over the range $\tau \in [1, 10]$ for all three of TRAINSET, DEVSET, and TESTSET. Q^{10} achieves 0.274371 bits/frame on DEVSET, which is 0.014200 bits/frame lower than the 0.288571 bits/frame achievable when only that target participant's speech activity is considered (not shown in the figure, but henceforth lowercase q^{10}).

In all subsequent experiments in this subsection, the conditioning context consists of $\mathbf{Q}_1^{10}$ — all 10 most recent frames of speech activity from both participants to the conversation — plus the τ most recent frames of one of several representations of pitch for the target participant. The first of these is just $\mathbf{P}$, as computed in Subsection 3.3. As can be seen in Figure 1 (where for notational convenience the lowercase "p" indicates target participant only), the most recent frame of pitch $\mathbf{P}_1^1$ by itself already provides an improvement over $\mathbf{Q}_1^{10}$ for TRAINSET. It appears that reductions in TRAINSET cross entropy rates begin to asymptote at $\tau = 3$ frames[5]. This indicates that the proposed model learns to exploit pitch for speech activity prediction, and that therefore recent pitch must be correlated with incipient speech activity in TRAINSET. The fact that the same trends are observed for DEVSET indicates that the correlations which the model learns on TRAINSET generalize to data unseen during model training. The model achieves a cross entropy rate minimum on DEVSET at $\tau = 3$ of 0.270831 bits/frame, which is 0.0035400 bits/frame lower than the best value for $\mathbf{Q}_1^{10}$ alone.

Absolute pitch, as represented by $\mathbf{P}$, is patently speaker-dependent; for the model to have successfully leveraged absolute pitch, it must be ignoring a significant portion of the variability observed in $\mathbf{P}$. To quantify this, an experiment was conducted which uses binary voicing $\mathbf{V}$ (instead of $\mathbf{P}$), whose elements $\mathbf{v}_k[n]$ are unity if the corresponding $\mathbf{p}_k[n]$ is non-NaN and zero otherwise. Denoted as "$\mathsf{Q}^{10} \cup \mathsf{v}^{\tau}$" in Figure 1, the curve exhibits a minimum on DEVSET at $\tau = 8$ of 0.271698, which 0.0026730 lower than for $\mathbf{Q}_1^{10}$ alone and represents 76% of the reduction observed for $\mathbf{P}$. This is suprisingly high and implies that the actual value of absolute pitch is not as relevant for prediction as is its (non-NaN) existence. Exposing the model to both $\mathbf{V}$ and $\mathbf{P}$ for the target participant (in addition to $\mathbf{Q}_1^{10}$), denoted "$\mathsf{Q}^{10} \cup \mathsf{v}^{\tau} \cup \mathsf{p}^{\tau}$" in Figure 1, is seen to lower the cross entropy rate to 0.270304 bits/frame at $\tau = 9$, by 0.000527 bits/frame. It is possible that the availability of $\mathbf{V}$ allows the model to focus on extracting information from frames in which absolute pitch is known to exist, and not waste its finite capacity on inferring this by itself.

Since, as expected, variability in absolute pitch $\mathbf{P}$ appears to present a problem for the model, an experiment was conducted which Z-normalizes $\mathbf{P}$ by each speaker's mean and standard deviation. These two quantities must be known a priori; assuming that they do not deviate from a speaker's conversation-specific statistics permits their estimation from each conversation separately. This leads to a new representation, $\mathbf{Z}$, whose elements $\mathbf{z}_k[n]$ are equal to $(\mathbf{p}_k[n] - \mu_P)/\sigma_P$ where $\mathbf{p}_k[n]$ is non-NaN, and NaN otherwise. The curve in Figure 1, denoted "$\mathsf{Q}^{10} \cup \mathsf{z}^{\tau}$", exhibits a DEVSET minimum of 0.270877 bits/frame at $\tau = 8$.

[4] For each experimental setting, a single randomly seeded model was trained.

[5] It should be noted that each model at τ, visually connected by a line to the point at $\tau - 1$, contains all of the features of that point. As a result, the curves can reasonably be expected to be monotonically decreasing or asymptotically flat. That they are not reflects the effect of random seeding and the fact that each point represents one model rather than an average over multiple, differently-seeded but otherwise-same, models.

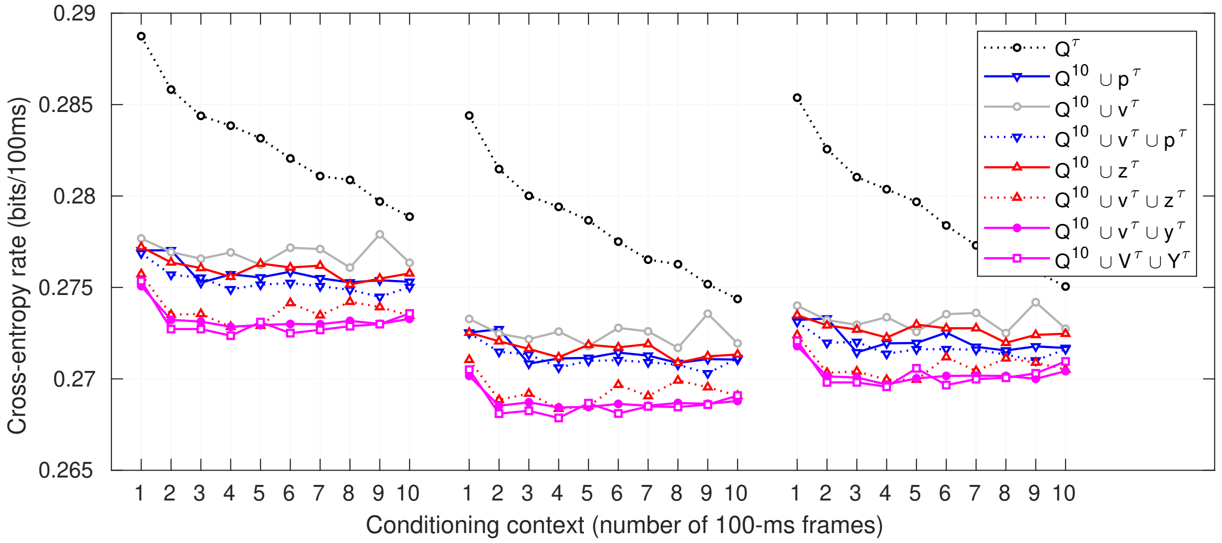

Figure 1: Cross entropy rate, along the y-axis in bits per 100-ms frame, for several representations of pitch on top of 10 frames of speech activity from both participants, as a function of the duration of the pitch history, along the x-axis in number of 100-ms frames. Rates are shown from left to right for TRAINSET, DEVSET, and TESTSET.

This is only negligibly different from the minimum of 0.270831 bits/frame achieved for absolute pitch (cf. the previously-discussed curve denoted "$\mathrm{Q}^{10} \cup \mathrm{p}^\tau$") at $\tau = 3$, and at first glance suggests the infelicity of Z-normalization. Closer inspection reveals that while Z-normalization usefully removes inter-speaker variability, it also brings values close to the speaker's mean close to zero, which makes them — from the model's point of view — indistinguishable from unvoiced frames. Exposing the model to both $\mathbf{V}$ and $\mathbf{Z}$ corrects this, and yields a cross entropy rate of 0.268366 at $\tau = 4$, as can be seen in Figure 1 for the curve denoted $\mathrm{Q}^{10} \cup \mathrm{v}^\tau \cup \mathrm{z}^\tau$. This is lower than the rate achieved by the $\mathbf{Q}^{10} \cup \mathbf{v}^\tau \cup \mathbf{p}^\tau$ curve by 0.0019380 bits/frame, almost 4 times more than the reduction observed when including $\mathbf{V}$ with $\mathbf{P}$.

Pitch is claimed to be perceived on a logarithmic scale; to explore whether log-pitch outperforms pitch on the speech activity prediction task, $\mathbf{L} \equiv \log_2 \mathbf{P}$ was formed. Its elements $\mathbf{l}_k[n]$ are equal to $\log_2 \mathbf{p}_k[n]$ when $\mathbf{p}_k[n]$ is non-NaN, and NaN otherwise. Z-normalizing $\mathbf{L}$ instead of $\mathbf{P}$ yields a new representation $\mathbf{Y}$, whose elements $\mathbf{y}_k[n]$ are equal to $(\mathbf{l}_k[n] - \mu_L)/\sigma_L$ if $\mathbf{l}_k[n]$ is non-NaN, and NaN otherwise. Denoted by the curve "$\mathrm{Q}^{10} \cup \mathrm{v}^\tau \cup \mathrm{y}^\tau$" in Figure 1, this representation yields a DEVSET cross entropy rate minimum of 0.268441 at $\tau = 4$. This is actually higher than the DEVSET minimum of the "$\mathrm{Q}^{10} \cup \mathrm{v}^\tau \cup \mathrm{z}^\tau$" curve, but it is lower for all values $\tau \neq 4$, and also smoother over the entire $\tau \in [1, 10]$ range.

The last experiment of this subsection builds on the logarithmic version, including voicing and z-normalized log-pitch not just for the target participant but also for their interlocutor. This is denoted in Figure 1 by "$\mathrm{Q}^{10} \cup \mathrm{V}^\tau \cup \mathrm{Y}^\tau$", and its minimum is reached at $\tau = 4$ with a value of 0.267864 bits/frame. It can be tentatively concluded that model sensitivity to the non-target participant's recent pitch history reduces average surprise, by the small amount of 0.000577 bits/frame.

4.2 History Duration

Experiments in the previous subsection show that recent pitch appears to be correlated with incipient speech activity, and that a predictor exposed to 10 frames of most-recent speech activity should also be exposed to at least 4 most-recent frames of voicing ($\mathbf{V}_1^4$) and Z-normalized log-pitch ($\mathbf{Y}_1^4$). Although it cannot be concluded that this particular representation is optimal, it is the most optimal representation from amongst those investigated for the Switchboard corpus. The experiments shown in Figure 2 aim to establish whether this is true even when much longer histories of speech activity are considered; (Laskowski and Shriberg, 2012) had shown that speech activity histories as long as 8 s (80 100-ms frames, compressed quasi-logarithmically) continue to improve predictions.

Figure 2 depicts the same "Q^τ" curve shown in Figure 1, but extends this to $\tau = 20$ 100-

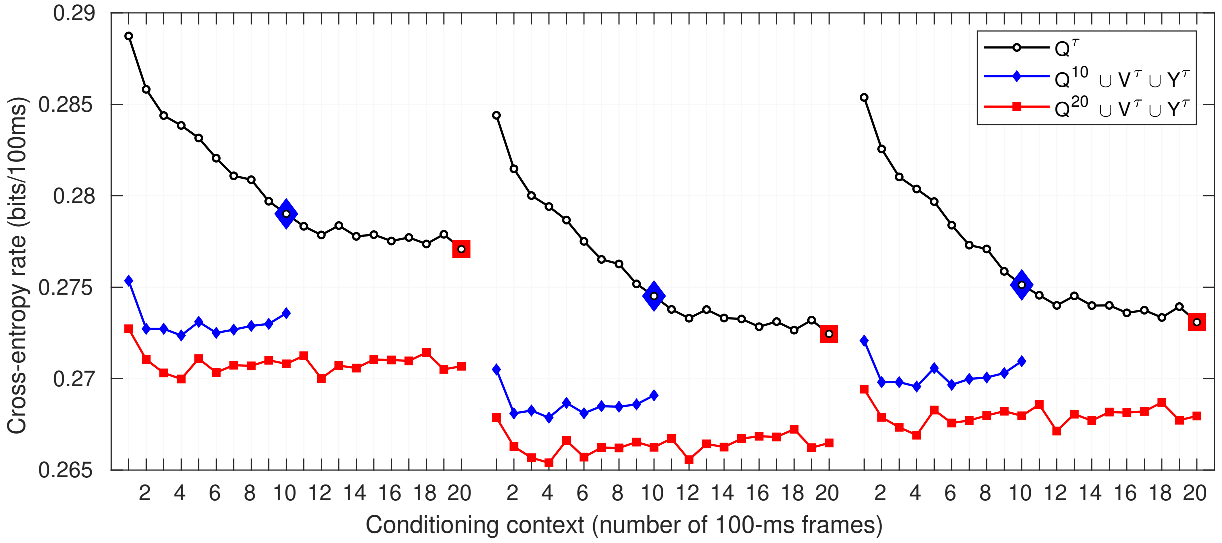

Figure 2: Cross entropy rate, along the y-axis in bits per 100-ms frame, for voicing ($\mathbf{V}$) and speaker-dependent Z-normalized log-pitch ($\mathbf{Y}$) on top of either 10 or 20 frames of speech activity from both participants (shown for reference with enlarged markers on the curve for $\mathbf{Q}$ alone), as a function of the duration of the pitch history, along the x-axis in number of 100-ms frames. Rates are shown from left to right for TRAINSET, DEVSET, and TESTSET. Lines connecting points are drawn for the purposes of visualizion.

ms frames of speech activity history. It can be seen, for both TRAINSET and DEVSET (as well as TESTSET), that the nearly-linear decrease in cross entropy rate as τ increases continues, albeit less steeply. Also shown in the figure is the same curve as "$Q^{10} \cup V^{\tau} \cup Y^{\tau}$", for which the DEVSET minimum can be found at $\tau = 4$. What is new in the figure is the curve denoted as "$Q^{20} \cup V^{\tau} \cup Y^{\tau}$", which depicts the impact of pitch when the speech activity history is 2 seconds rather than 1 second long. As can be seen, this third curve exhibits its DEVSET minimum also at $\tau = 4$. A system trained on $\mathbf{Q}_1^{10} \cup \mathbf{V}_1^4 \cup \mathbf{Y}_1^4$ reduces the cross entropy rate of a system trained on $\mathbf{Q}_1^{10}$ alone by $0.274515 - 0.267864 = 0.0066510$ bits/frame; one that is trained on $\mathbf{Q}_1^{20} \cup \mathbf{V}_1^4 \cup \mathbf{Y}_1^4$ exhibits a reduction over a system trained on $\mathbf{Q}_1^{20}$ alone by $0.272448 - 0.265396 = 0.0070520$ bits/frame. This is not only a larger reduction in absolute terms, it appears even larger relative to the speech-only baseline. It suggests that the usefulness of the most recent 400 ms of pitch grows as the duration of speech activity history increases.

4.3 Model Complexity and Training

A final suite of experiments was conducted in order to shed light on potential under-training or over-fitting of the model, given the fixed size of TRAINSET. The representation identified at the end of Subsection 4.1 was used, namely $\mathbf{Q}_1^{10} \cup \mathbf{V}_1^4 \cup \mathbf{Y}_1^4$; there, the model consisted of 8 units in its hidden layer and its training consisted of 1000 iterations of SCG descent. Figure 3 compares cross-entropy rates when the number of training iterations and the number of hidden units are varied in $\{1000, 2000, 3000, 4000\}$ and $\{8, 16, 32, 64\}$, respectively. Note that these numbers of hidden units correspond to 305, 609, 1217, and 2433 free parameters, given an input representation dimensionality of 36.

As can be seen in the figure, extending the training regimen to 2000 iterations is clearly beneficial; extending it further to 3000 iterations yields only negligibly lower DEVSET cross entropy rates. Increasing the model complexity from 8 to 64 hidden units is also beneficial, but on DEVSET the improvement from 32 to 64 units is much smaller than on TRAINSET, indicating not-yet overfitting but getting close. The DEVSET cross entropy rate for 64 units and 4000 iterations is already higher than that for 64 units and 3000 iterations. Note that there is no evidence that more than 400 ms of pitch might benefit any of these larger systems.

5 Discussion

5.1 Generalization

The models presented in this article have all been trained using TRAINSET alone; model selection has been conducted using cross entropy rate min-

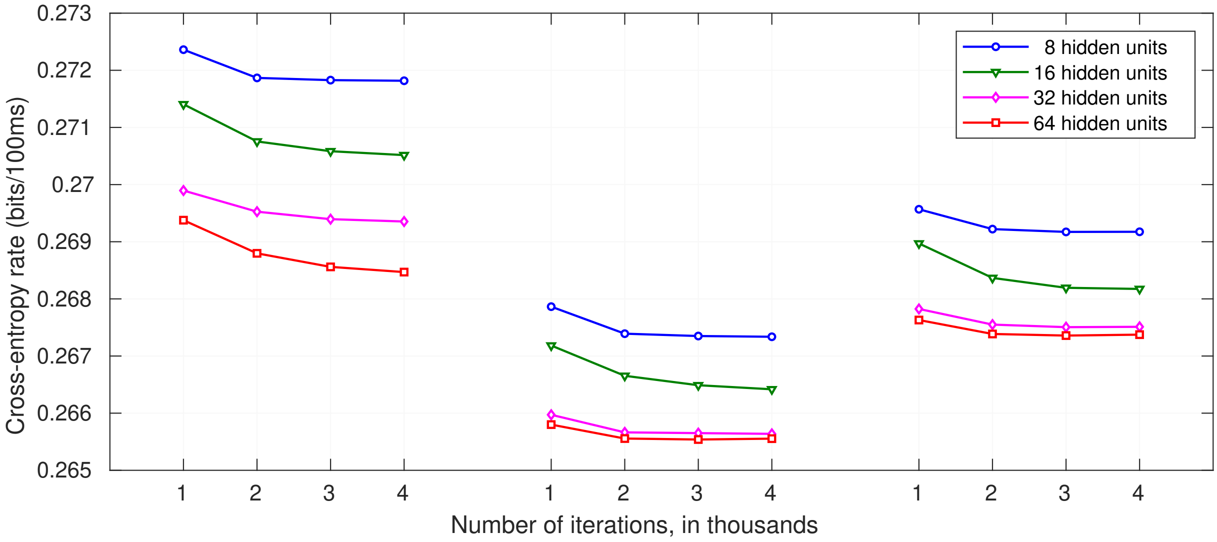

Figure 3: Cross entropy rate, along the y-axis in bits per 100-ms frame, for four models differing in the number H of hidden units and using 4 100-ms frames of voicing ($\mathbf{V}$) and speaker-dependent Z-normalized log-pitch ($\mathbf{Y}$) on top of 10 100-ms frames of speech activity from both participants, as a function of the number of iterations of SCG training, along the x-axis in thousands. Rates are shown from left to right for TRAINSET, DEVSET, and TESTSET. Lines connecting points are drawn for the purposes of visualizion.

Feature Set	H	I	$\mathcal{X}$
$\mathbf{Q}_1$	8	1000	0.285379
$\mathbf{Q}_1^{10}$	8	1000	0.275052
$\mathbf{Q}_1^{10} \cup \mathbf{V}_{tar,1}^{\tau=8}$	8	1000	0.272502
$\mathbf{Q}_1^{10} \cup \mathbf{P}_{tar,1}^{\tau=3}$	8	1000	0.271450
$\mathbf{Q}_1^{10} \cup \mathbf{V}_{tar,1}^{\tau=9} \cup \mathbf{P}_{tar,1}^{\tau=9}$	8	1000	0.271006
$\mathbf{Q}_1^{10} \cup \mathbf{V}_{tar,1}^{\tau=9} \cup \mathbf{Z}_{tar,1}^{\tau=4}$	8	1000	0.269953
$\mathbf{Q}_1^{10} \cup \mathbf{V}_{tar,1}^{\tau=9} \cup \mathbf{Y}_{tar,1}^{\tau=4}$	8	1000	0.269690
$\mathbf{Q}_1^{10} \cup \mathbf{V}_1^{\tau=9} \cup \mathbf{Y}_1^{\tau=4}$	8	1000	0.269568
$\mathbf{Q}_1^{10} \cup \mathbf{V}_1^{\tau=9} \cup \mathbf{Y}_1^{\tau=4}$	64	1000	0.267630
$\mathbf{Q}_1^{10} \cup \mathbf{V}_1^{\tau=9} \cup \mathbf{Y}_1^{\tau=4}$	64	3000	0.267358

Table 1: Cross entropy rates $\mathcal{X}$ in bits per 100-ms frame, obtained for TESTSET using several representations of pitch, numbers H of hidden units, and numbers I of training iterations. All models trained on TRAINSET, and model selection (over τ, H, and/or I as applicable) performed using DEVSET.

imization on DEVSET. TESTSET has been left untouched, and therefore presents a suitable candidate set for characterizing how the proposed framework generalizes to *completely* unseen data. Table 1 summarizes these achievements, from the right-hand-side of Figures 1 and 2.

As can be seen, the absolute reduction in cross entropy rate due to the inclusion of pitch information (in the form of voicing and Z-normalized log-pitch) is $0.275052 - 0.267358 = 0.0076940$ bits/frame. This magnitude represents approximately 75% of the reduction observed when pitch information is excluded and the speech activity context is increased from 1 frame to 10 frames ($0.285379 - 0.275052 = 0.010327$ bits/frame, ie. rows 1 and 2 in the table). All trends observed for TESTSET in Figures 1, 2, and 3 are nearly identical to those observed for DEVSET.

5.2 Normalization

That the prediction of speech activity can successfully make use of approximately 8 s of most-recent speech activity history (Laskowski and Shriberg, 2012), but of only 400 ms of most-recent pitch history, is surprising and somewhat deflating. However, it is important to note that the optimal representation of pitch was determined to involve Z-normalization, for which the conversation-side mean and standard deviation were assumed to be known a priori. In reality, these statistics would need to be accumulated from the start of each conversation, up to and including the $(n-1)$th frame. It is also possible that estimation of these statistics should favor the recent past, yielding local Z-normalization statistics which themselves evolve over time. This is currently under investigation.

5.3 Reproducibility

The experiments presented in this article number just shy of 150; each experiment took approximately 6 hours to run on a hyper-threaded 6-core Intel Xeon E5645 2.40GHz machine, running Debian Linux 3.16. The complete experiment suite, including all source and intermediate Switchboard Corpus data, are available at `www.cs.cmu.edu/~kornel/software/stt.html`.

5.4 Potential Impact

For Switchboard conversations, the proposed framework has demonstrated that attentiveness to the pitch trajectories of both conversation sides reduces the average surprise of incipient side-attributed speech activity. It appears that it suffices for the considered pitch trajectories to be quite short (400 ms). The Switchboard corpus thereby provides sufficient proof that the proposed framework is capable of yielding findings such as these, in cases in which only the actual speech activity is available and for which pitch can be automatically measured. The framework is agnostic to the much more contentious attempts to define and annotate what a turn is, and not reliant on additional turn-landmark or pitch-trajectory annotation.

The direct impact of this work is that it enables the automated analysis — with regard to the role of pitch in turn-taking — of large corpora which would otherwise be intractable to analyze in their entirety. Due to its quantitative nature, the framework enables direct comparisons between corpora which differ in arguably important ways, such as language, dialect, or domain.

Furthermore, an indirect impact of the findings of which the proposed framework is capable is that such findings may inform automated speech processing systems operating under specific language, dialect, or domain conditions, for example mixed-initiative dialog systems. Knowledge of how such conditions affect the interplay between pitch and turn-taking would enhance the naturalness and flexibility of those systems.

6 Conclusions

Pitch has long been held as an important signalling channel when planning and deploying speech in conversation, and myriad studies have been undertaken to determine the extent to which it actually plays this role. Unfortunately, these studies have required considerable human investment in data preparation and analysis, and have therefore often been limited to a handful of specific conversational contexts. This has made it difficult to compare and contrast, in a quantitative way, the role played by pitch in turn-taking as a function of language, dialect, domain, channel, other-party familiarity, etc.

The framework proposed in this article addresses these limitations, by enabling a nearly-automatic quantitative characterization of the role of pitch throughout an entire conversation, requiring only the raw signal and speech activity references. Although the latter may require prior manual transcription of the lexical content (followed by forced alignment), this is far easier than manually annotating turn landmarks or pitch trajectories, and is often already available for a corpus under study. The framework is adaptible to the role-in-turn-taking analysis of any feature which can be measured from the raw signal.

This article has evaluated the proposed framework by answering three specific questions regarding the role of pitch in turn-taking, in the Switchboard corpus. First, the presented evidence suggests that pitch can be leveraged to reduce the average surprise of incipient speech. Its inclusion, on top of a conditioning context containing 1 second of speech activity from both dialogue parties, yields a cross entropy reduction of 0.014200 bits per 100 ms; this is approximately half as much as is gained by including the non-target participant's 1-second of speech activity, over just the target participant's, in the first place. Second, the optimal representation of pitch appears to be Z-normalized log-pitch, together with the binary indicator variable of voicing; at least in part, the role of the latter is to differentiate between unvoiced frames and voiced mean-log-pitch frames. Finally, experiments indicate that the dynamic pitch trajectory information which is useful for speech activity prediction is limited to the most recent 400 ms; pitch trajectory information less recent than that is necessary only to provide static Z-normalization statistics. Furthermore, the reduction in average surprise appears to be a function of the duration of the considered speech activity history; the longer the speech activity history, the more valuable do those most recent 400 ms of pitch seem to be.

Acknowledgments

This work was funded in part by the Stiftelsen Marcus och Amalia Wallenbergs Minnesfond project MAW 2017.0034, *Hidden events in turn-taking*.

References

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

S. Bögels and F. Torreira. 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

J. Caspers. 2003. Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics*, 31(2):251–276.

N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone. 1998. Resegmentation of SWITCHBOARD. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP1998)*, pages unnumbered, paper 0685, Sydney, Australia.

S. Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

J. Edlund and M. Heldner. 2005. Exploring prosody in interaction control. *Phonetica*, 62(2-4):215–226.

C. Ford and S. Thompson. 1996. *Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns*, chapter 3. Cambridge University Press, Cambridge MA, USA.

J. Godfrey and E. Hollimann. 1997. *Switchboard-1 Release 2*. Catalog Number LDC97S62, Linguistic Data Consortium, Philadelphia PA, USA.

A. Gravano and J. Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25(3):601–634.

M. Heldner, J. Edlund, K. Laskowski, and A. Pelcé. 2009. *Prosodic features in the vicinity of silences and overlaps*, pages 95–105. Peter Lang, Frankfurt am Main, Germany.

M. Heldner and M. Włodarczak. 2015. *Pitch slope and end point as turn-taking cues in Swedish*, pages 10–15. Glasgow, Scotland.

A. Hjalmarsson. 2011. The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1):23–35.

G. Jefferson. 1984. *Transcript notation*, pages ix–xvi. Cambridge University Press, Cambridge MA, USA.

J. Kane, I. Yanushevskaya, C. de Looze, B. Vaughan, and A. Ní Chasaide. 2014. *Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions*, pages 333–337. Singapore.

H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech*, 41(3–4):295–321.

K. Laskowski. 2012. Exploiting loudness dynamics in stochastic model of turn-taking. In *Proceedings of the 4th IEEE Workshop on Spoken Language Technology (SLT2012)*, pages 79–84, Miami FL, USA.

K. Laskowski. 2015. Auto-imputing radial basis functions for neural network turn-taking models. In *Proceedings of the 15th Annual Conference of the International Speech Communcations Association (INTERSPEECH2015)*, pages 1820–1824, Dresden, Germany.

K. Laskowski, M. Heldner, and J. Edlund. 2009. *Exploring the prosody of floor mechanisms in English using the fundamental frequency variation spectrum*, pages 2539–2543. Glasgow, Scotland.

K. Laskowski and E. Shriberg. 2012. Corpus-independent history compression for stochastic turn-taking models. In *Proceedings of the 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2012)*, pages 4937–4940, Kyoto, Japan.

J. Local, J. Kelly, and W. Wells. 1986. Towards a phonology for conversation: Turn-taking in Tyneside English. *Journal of Linguistics*, 22(2):411–437.

J. Local and G. Walker. 2012. How phonetic features project more talk. *Journal of the International Phonetic Association*, 42(3):255–280.

M. Møller. 1993. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533.

M. Selting. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics*, 6:357–388.

K. Sjölander. 2001. Snack sound toolkit 2.2.10. `http://www.speech.kth.se/snack/`.

G. Walker. 2017. Pitch and the projection of more talk. *Research on Language and Social Interaction*, 50(2):206–225.

A. Wennerstrom and A. F. Siegel. 2003. Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, 36(2):77–107.

I. Yanushevskaya, J. Kane, C. de Looze, and A. Ní Chasaide. 2014. pages 959–963. Dublin, Ireland.

M. Zellers. 2013. *Pitch and lengthening as cues to turn transition in Swedish*, pages 248–252. Lyon, France.

M. Zellers. 2017. Prosodic variation and segmental reduction and their roles in cuing turn transition in swedish. *Language and Speech*, 60(3):454–478.