AN INSTANTANEOUS VECTOR REPRESENTATION OF DELTA PITCH FOR SPEAKER-CHANGE PREDICTION IN CONVERSATIONAL DIALOGUE SYSTEMS

Kornel Laskowski[1]*, Jens Edlund*[2]*, and Mattias Heldner*[2]

[1] interACT, Carnegie Mellon University, Pittsburgh PA, USA
[2] Centre for Speech Technology, KTH, Stockholm, Sweden
kornel@cs.cmu.edu

ABSTRACT

As spoken dialogue systems become deployed in increasingly complex domains, they face rising demands on the naturalness of interaction. We focus on system responsiveness, aiming to mimic human-like dialogue flow control by predicting speaker changes as observed in real human-human conversations. We derive an instantaneous vector representation of pitch variation and show that it is amenable to standard acoustic modeling techniques. Using a small amount of automatically labeled data, we train models which significantly outperform current state-of-the-art pause-only systems, and replicate to within 1% absolute the performance of our previously published hand-crafted baseline. The new system additionally offers scope for run-time control over the precision or recall of locations at which to speak.

Index Terms— Speech communication, User interfaces, Speech processing, Frequency domain analysis, Signal representation.

1. INTRODUCTION

As spoken dialogue systems (SDSs) become deployed in increasingly complex domains, such as tutoring, entertainment and games, they face rising demands on the naturalness of interaction. A *conversational* SDS, which aims to be "human enough that we respond to it as we respond to another human" [1], needs to be flexible, robust, and responsive. Although all SDSs must identify places where they can legitimately begin talking without interrupting, this is especially true for systems which strive to mimic human behavior. A large number of inappropriately late responses can ruin the illusion of having a conversation with the system. However, conversational speech also contains many long ($\geq$0.5s) within-utterance pauses, where a careless SDS may inappropriately barge in.

Current state-of-the-art SDSs identify suitable places to speak with the help of an end-of-utterance (EOU) detector. This component relies on speech activity detection (SAD), which marshals spoken input into contiguous intervals of speech with internal pauses no longer than a predefined threshold; we will refer to these intervals as *talkspurts* [2]. Candidate EOUs are considered only at end-of-talkspurt (EOT) events. In the majority of SDSs currently in use, this determination is based exclusively on the duration of the post-EOT pause. Such systems tend to favor politeness over a high number of interruptions by choosing long pauses (1–2s), rendering the conversation less responsive than with a human interlocutor.

Research which has addressed this issue has relied on SAD post-processing for faster EOU prediction, using automatic speech recognition (ASR) output [3, 4], models of prosody [5], or a combination of the two [6]. With the exception of our own work [5], these studies have focused on improving human-computer speech by studying *existing, pause-governed* human-computer interaction; it is therefore unclear how well-suited they are for improving models of more human-like conversational speech.

Data Set	Duration (mn:ss)	Dialogue role g		
		speakers	# EOTs	# SCs
DEVSET	77:40	F4,F5,M2,M3	480	222
EVALSET	60:39	F1,F2,F3,M1	317	149

Table 1. Size, speakers, number of end-of-talkspurt (EOT) and speaker change (SC) events for the speaker in role g in our datasets.

In this work, we explore when humans choose to speak, in highly interactive human-human dialogues (described in Section 2). We characterize an inexpensive, automatic means [5] of assigning labels, based on human interlocutor behavior, in Section 3. In Section 4, we evaluate these labels with respect to several baselines. Using the labels as targets, we train acoustic models of prosodic variation to predict, in unseen data, whether humans would or would not begin speaking. To achieve this, we propose a vector representation of delta pitch, or variation in fundamental frequency (F0), in Section 5, which differs from current scalar-valued representations in that it is: (1) *instantaneous*, not relying on adjacent frames; (2) *continuous*, defined for all time; (3) distributed; and (4) potentially *sparse*, suitable for the application of standard acoustic modeling techniques. We present validating experiments in Section 6.

2. HUMAN-HUMAN DIALOGUE CORPUS

For our study, we use Map Task [7] dialogues, representatative of spontaneous collaborative speech. A Map Task dialogue has two participants: a *giver* (g), providing instructions, and a *follower* (f). The task is for the giver to describe a route indicated on his or her map to the follower. The differences between this domain and less interactive, currently studied human-computer domains such as ATIS are significant; for example, [8] cites a range for disfluencies (which include pauses) of 0.8–2.1% of words for human-computer interaction and 5.5–7.3% of words for human-human interaction, with the higher numbers representative of Map Task-like domains.

The specific corpus used here is the Swedish Map Task Corpus [9]. We divided this data into a development set (DEVSET) and an evaluation set (EVALSET) which are disjoint in speakers (cf. Table 1). The division was chosen to enable comparison with our published work [5] on the EVALSET. For the same reason, we use EOTs only from the speaker in role g, extracted using the current version of the SAD (with a minimum pause of 300 ms) used in that work.

3. LABELING FRAMEWORK

Manual annotation of human-human dialogue is labor-intensive and therefore expensive. Since large amounts of data are likely to be needed to train suitable models of human-human behaviour, there is scope for alternatives to human labeling. In previous work [5], we used the presence of *observed* speaker change as a measure of appropriateness. In this section, we explain these labels in detail.

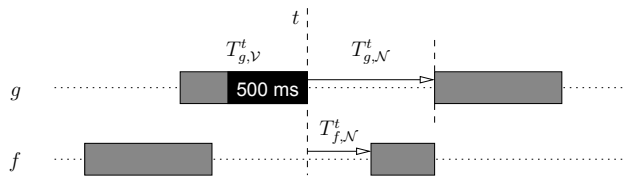

Fig. 1. g's and f's talkspurts in the vicinity of an EOT in g's speech at instant t; time shown from left to right. Symbols as in Section 3.

In considering whether a particular talkspurt of speaker g, terminating at time t, is likely to be followed by a speaker change, we inspect how speaker f behaved under the circumstances, shown in Fig. 1. The pause between the currently terminating talkspurt and g's next talkspurt has duration $T^t_{g,\mathcal{N}}$. We do not consider overlapped vocalization. f produces his or her next talkspurt at $t+T^t_{f,\mathcal{N}}$. Given only this characterization, we can inspect whether, following time t, the next talkspurt was produced by f or by g. We refer to the former as *a speaker change* (SC), and the latter as *not a speaker change* ($\neg$SC). Therefore, the label assigned to the EOT at time t is

$$L_t \quad = \quad \begin{cases} \text{SC} & \text{if } T^t_{f,\mathcal{N}} - T^t_{g,\mathcal{N}} < 0 \\ \neg\text{SC}, & \text{otherwise} \end{cases} \qquad (1)$$

In this work, online estimation of appropriateness to vocalize at time t by the system consists of predicting the value of L_t given a prosodic description of the last 500 ms of speech terminating at time t (shown in black in Fig. 1).

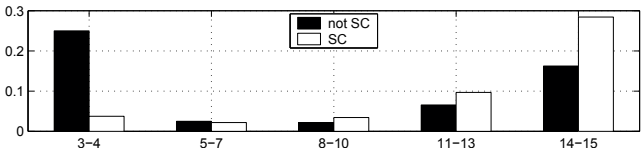

Fig. 2. Normalized distribution of human-annotated appropriateness labels over the automatically produced SC and $\neg$SC labels; the $x-$axis shows summed appropriateness votes from three labelers.

4. BASELINE

Although prosody can sometimes signal whether an EOT is appropriate for SC, the appropriateness of an SC does not guarantee its occurrence. To evaluate the relationship between these two factors, we compared the automatic L_t labels with human assessment made independently by three judges. Presented with 1.5s of audio preceding every EVALSET EOT, in random order, each judge was asked whether the EOT was an appropriate place to speak on a five-point Likert scale. Inter-labeler consistency was high; α for the five-point scale was 0.95, and pairwise κ for the binary decision *appropriate* vs *inappropriate* (i.e. pooling 1-3 and 4-5 on the scale) was 0.70–0.88. The sum of the three Likert scores was used as a descriptive measure of appropriateness; we show the distribution of scores over automatic L_t in Fig. 2. There is a clear correlation between high appropriateness and an actual SC, and a very strong correlation between $\neg$SC and those EOTs judged as highly inappropriate.

Fig. 3 shows receiver operating characteristic (ROC) curves for SC/$\neg$SC discrimination for three baseline methods. First, we produced a curve by varying a threshold for the sum of the human appropriateness labels. The algorithm predictably overgenerates, a characteristic of dialogue types in which SCs are optional; the highest dot, representing the most aggressive system with a threshold of 4, rejects only unanimously inappropriate EOTs, but still incurs a false positive (FP) rate of 50%. Dialogue types in which the most appropriate places for SCs are only actual SCs, such as question-answering, tend to show a low FP rate. The ROC curve in [6], for instance, indicates a FP rate of $<$13% if every EOT is selected.

The second curve in Figure 3, for a system relying exclusively on pause length, was computed following [6]. Also shown are the results for our baseline hand-crafted automatic system in [5]. They represent the performance of an aggressive and a non-agressive version of the system. Note that the baseline system can only be operated at the two levels of aggressiveness shown.

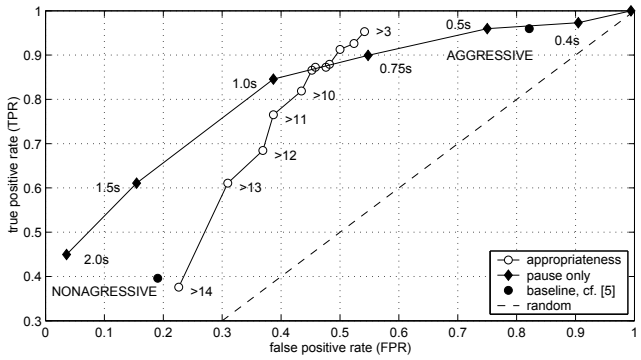

Fig. 3. ROC curves for the baselines described in Section 4.

It should be noted that the three baselines represent quite different temporal situations. The pause threshold system responds after 0.3—2.0 s depending on which part of the curve we examine, the system used in [5] responds in 0.3 s in both cases, and the human annotators had no knowledge of the pause length following the stimuli they heard, so they can be said to respond instantly. When the pause threshold system incurs fewer FPs, it is significantly slower.

5. THE DELTA PITCH REPRESENTATION

Instantaneous variation in pitch is normally computed by determining a single scalar, the fundamental frequency (F0), at two temporally adjacent instants and forming their difference. F0 represents the frequency of the first harmonic in a spectral representation of a frame of audio, and is undefined for signals without harmonic structure. In the context of speech processing applications, we view the localization of the first harmonic, and the subsequent differencing of two adjacent estimates, as a case of suboptimal feature compression and premature inference, since the goal of such applications is not

the accurate estimate of pitch. In particular, we would like to leverage the fact that *all* harmonics are spaced equally in each of the two adjacent frames, and use *every* element of a spectral representation to yield a representation of the F0 delta.

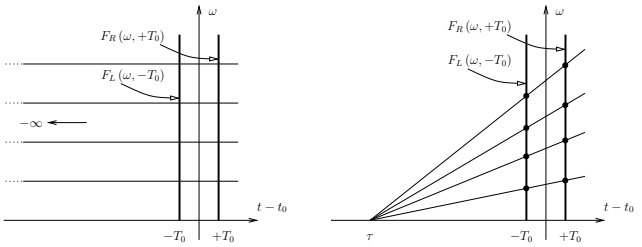

Fig. 4. The standard dot-product shown as an orthonormal projection onto a point at infinity, and the proposed vanishing-point product, which generalizes to the former when $\tau \rightarrow \infty$.

To this end, we propose a vector-valued representation of pitch variation, inspired by *vanishing-point perspective*, a technique used in architecural drawing and grounded in projective geometry. While the standard inner product between two vectors can be viewed as the summation of pair-wise products with pairs selected by orthonormal projection onto a point at infinity, the proposed *vanishing-point product* induces a 1-point perspective projection (see Fig. 4). In effect, frequency dilation or compression of one or both of the frequency representations is controlled by the distance τ. The proposed vector-valued representation of pitch variation is the vanishing-point dot-product, evaluated over a continuum of τ.

In computing the vanishing-point product $g^{\tau}_{t_0}(\tau)$ at time t, we consider the short-time frequency representation of the left-half and the right-half portion of a single analysis window centered at time t_0. We refer to these as $F_{t_0,L}\left(e^{j\omega}\right)$ and $F_{t_0,R}\left(e^{j\omega}\right)$, respectively (but drop the t_0 subscript in the ensuing derivation for clarity); they are obtained using two asymmetrical windows which are mirror-images of each other, as shown in Fig. 5.

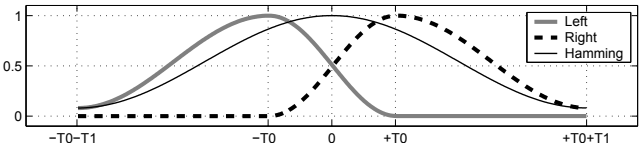

Fig. 5. Left and right windows used for the computation of F_L and F_R, respectively, consisting of asymmetrical Hamming window halves. T_0 is 4 ms, and T_1 is 12 ms, for a full analysis window width of 32 ms. A 32 ms Hamming window is shown for comparison.

The vanishing-point projection metaphor allows for a direct derivation of $g^{\tau}(\tau)$ from Fig. 6 (the mirror image of the figure, for $\tau > +T_0$, is not shown due to space constraints) to yield:

$$g^{\tau}(\tau) \quad (2)$$
$$= \begin{cases} \int_{-f_s/2}^{+f_s/2} F_L\left(\left(\frac{-\tau-T_0}{-\tau+T_0}\right)f\right) F_R^*(f)\, df & \tau < -T_0 \\ \int_{-f_s/2}^{+f_s/2} F_L(f)\, F_R^*\left(\left(\frac{+\tau-T_0}{+\tau+T_0}\right)f\right) df & \tau > +T_0 \end{cases}$$

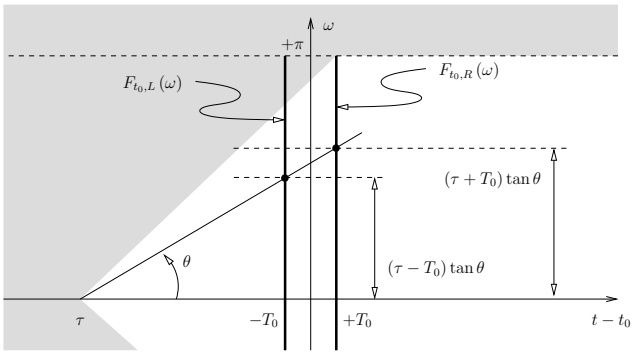

Fig. 6. Computation of $g^{\tau}(\tau)$ for $\tau < -T_0$; F_L and F_R are at $-T_0$ and $+T_0$, respectively. $\theta_0 = \arctan f_s/2\,(|\tau| + T_0)$.

We then define a conformal mapping of τ onto

$$\rho = \begin{cases} -\log_2\left(\frac{-\tau-T_0}{-\tau+T_0}\right) & \tau < -T_0 \\ +\log_2\left(\frac{+\tau-T_0}{+\tau+T_0}\right) & \tau > +T_0 \end{cases} \quad (3)$$

It can be observed that ρ and τ are conveniently of opposite sign (at $\arg\max_{\tau} g^{\tau}(\tau) < -T_0$, F0 is increasing, and vice versa). This permits rewriting Eq. 2 as

$$g^{\rho}(\rho) = \begin{cases} \int_{-f_s/2}^{+f_s/2} F_L(f)\, F_R^*\left(2^{+\rho} f\right) df & \rho < 0 \\ \int_{-f_s/2}^{+f_s/2} F_L\left(2^{-\rho} f\right) F_R^*(f)\, df & \rho \geq 0 \end{cases} \quad (4)$$

In practice, we compute Eq. 4 using magnitude spectra, rather than complex spectra. $|F_L|$ and $|F_R|$ represent discrete transforms (of length $N = 512$, over [-256,255]), in general necessitating interpolation, $|\tilde{F}\left(2^{\pm\rho}k\right)| = \beta\,|F\left[\lfloor 2^{\pm\rho}k \rfloor\right]| + (1-\beta)\,|F\left[\lceil 2^{\pm\rho}k \rceil\right]|$, where $\beta = \left|\lceil 2^{\pm\rho}k \rceil - 2^{\pm\rho}k\right|$. We sample the transform at the equispaced locations $\rho = 4r/N$, $-N/2 \leq r < N/2$, representing a range of $[-2, +2)$ octaves, to yield

$$g^{\rho}[r] = \begin{cases} \sum_{k=-N/2+1}^{N/2} |\tilde{F}_L\left(2^{-4r/N}k\right)|\,|F_R^*[k]| & r \geq 0 \\ \sum_{k=-N/2+1}^{N/2} |F_L[k]|\,|\tilde{F}_R^*\left(2^{+4r/N}k\right)| & r < 0 \end{cases} \quad (5)$$

For subsequent modeling, we normalize Eq. 5 by the square root of $\sum|F_L|^2 \cdot \sum|F_R|^2$, with either $|F_L|$ or $|F_R|$ dilated as in Eq. 5, to yield an energy-independent vector representation.

6. EXPERIMENTS AND DISCUSSION

We now present several experiments in which, given the 500 ms preceding an EOT at t, we attempt to predict whether L_t is SC or $\neg$SC. $g^{\rho}[r]$ is computed every 8 ms. Each $g^{\rho}[r]$ is passed through a filterbank which attempts to capture meaningful prosodic variation. Our previous work [5] has shown falling (and low) intonation patterns before pauses to be a predictor of SC and flat intonation patterns (in the middle of a speaker's range) to be a strong predictor of $\neg$SC. The filterbank therefore contains a conservative trapezoidal filter for perceptually "flat" pitch [10], with a half-max span of $(-0.19, +0.19)$ semitones$/2T_0$, and two trapezoidal filters for perceptually "changing" pitch, with half-max spans of $(-0.66, -0.09)$ and $(+0.09, +0.66)$ semitones$/2T_0$. We also include two rectangular filters with spans of $(-2, -1)$ and $(+1, +2)$ octaves$/2T_0$, as we

have observed that unvoiced frames have flat rather than decaying tails. This filterbank reduces the input space to 5 scalars per frame.

We model each of the SC and ¬SC classes with a fully-connected hidden Markov model (HMM) of 4 states with one Gaussian per state, using Kevin Murphy's HMM Toolbox[1]. For development, the models are trained on the DEVSET using four-fold leave-one-out validation. We repeat this ten times, and average over the log-likelihoods obtained, before applying log-likelihood ratio selection. For final evaluation, the models are trained on the whole DEVSET, and applied to the EOTs in the EVALSET.

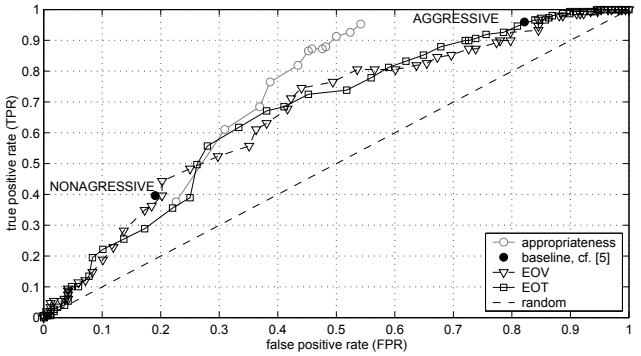

Fig. 7. ROC curves for systems EOT and EOV (cf. Section 6).

In Fig. 7, we report the performance of the described system, labeled "EOT", on the EVALSET. For contrast, we trained a second system ("EOV" in the figure) not on the last 500 ms preceeding each EOT, but on the 500 ms preceeding the last voiced frame prior to each EOT, as we had noted that some of our automatically detected EOTs contained a significant post-speech interval of exhalation. Voicing was determined using the Snack Sound Toolkit[2]. Fig. 7 also shows the appropriateness and automatic baselines.

It is apparent from the figure that our new speaker change prediction algorithm is significantly better than random, everywhere; this appears to validate our novel representation of delta pitch, our filterbank design, as well as the rudimentary HMM structure we selected. Furthermore, the "EOT" system replicates the performance of our AGGRESSIVE baseline and the "EOV" system comes within 1%abs of the NONAGGRESSIVE baseline, even though both baselines relied not only on pitch change but also location within pitch range, which neither new system accounts for. Furthermore, the "EOV" system performs close to human appropriateness judgment in the high precision range. Finally, similarity between the "EOT" and "EOV" curves shows that "EOT" performance is quite robust to SAD errors.

The experiments presented here validate the applicability of our approach on a single task, using one corpus in one language, and one side of the dialogue. A cross-domain/corpus evaluation is needed to establish the extent to which the approach generalizes. Our immediate future goal will be the manual inspection of what the learned models actually capture, and whether that corroborates existing studies of human speech. The current work also points to several new avenues of inquery. Due to the low cost of labeling, larger amounts of data can be used to augment the robustness of the current models. In conjunction, experiments like those presented can be used to optimize both model topologies and filterbank structures, in order to more accurately capture what humans may perceive as salient.

[1] http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

[2] http://www.speech.kth.se/snack/

7. CONCLUSIONS

We have derived a continuous vector representation of instantaneous variation in fundamental frequency, and showed that it is compatible with standard acoustic modeling techniques as used in SAD and ASR. In spite of label mismatch, we have shown that when trained only towards what actually occurs in human-human conversations, our models successfully discard a high number of locations to speak which are judged inappropriate by humans, in cases matching human performance. We have replicated, to within 1%, the performance of a hand-crafted baseline, without the use of pitch range information, and with the implicit benefit of run-time control of the recall or precision rates; at the studied delay of only 300 ms, our system significantly outperforms current state-of-the-art pause-only systems. Finally, as its training completely obviates the need for human labeling, the system has high potential for performance improvement due to larger training corpora, and important scope for data-driven prosodic model construction in dialogue systems and other domains.

8. ACKNOWLEDGMENTS

We thank Joakim Gustafson for help with the manual annotation; Pétur Helgason for access to the Swedish Map Task corpus; Tanja Schultz and Rolf Carlson for encouraging this collaboration; and Anton Batliner, Rob Malkin, Rich Stern, Ashish Venugopal, and Matthias Wölfel for several occasions of discussion. This work was funded in part by the Swedish Research Council (VR) project, 2006-2172, and in part by DARPA under contract HR0011-06-2-001.

9. REFERENCES

[1] J. Cassell, "Body language: lessons from the near-human," in *Genesis Redux*, J. Riskin, Ed. University of Chicago Press, Chicago, IL, USA, 2007.

[2] A.C. Norwine and O.J. Murphy, "Characteristic time intervals in telephonic conversation," *The Bell System Technical Journal*, vol. 17, pp. 281–291, 1938.

[3] L. Bell, J. Boye, and J. Gustafson, "Real-time handling of fragmented utterances," in *Proc. NAACL Workshop on Adaptation in Dialogue Systems*, pp. 2–8. Pittsburgh, PA, USA, 2001.

[4] G. Skantze and J. Edlund, "Robust interpretation in the Higgins spoken dialogue system," in *ITRW on Robustness Issues in Conversational Interaction*, Norwich, UK, 2004.

[5] J. Edlund and M. Heldner, "Exploring prosody in interaction control," *Phonetica*, vol. 62, pp. 215–226, 2005.

[6] 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 2002*, Denver CO, USA, 2002, vol. 3, pp. 2061–2064.

[7] A.H. Anderson et al., "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, no. 4, pp. 83–97, 1991.

[8] S.L. Oviatt, "Multimodal interfaces for dynamic interactive maps," in *Proc. CHI '96*, New York, NY, 1996, pp. 95–102.

[9] P. Helgason, "SMTC - a Swedish Map Task corpus," in *Working Papers 52: Proc. Fonetik 2006*, Lund, 2006, pp. 57–60.

[10] J. 't Hart, R. Collier, and A. Cohen, *A perceptual study of intonation*, Cambridge University Press, Cambridge, 1990.