EXPLORING THE PROSODY OF FLOOR MECHANISMS IN ENGLISH USING THE FUNDAMENTAL FREQUENCY VARIATION SPECTRUM

Kornel Laskowski [1,2], *Mattias Heldner* [3] *and Jens Edlund* [3]

[1] Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
[2] Institut für Anthropomatik, Universität Karlsruhe, Karlsruhe, Germany
[3] KTH Speech, Music and Hearing, Stockholm, Sweden
kornel@cs.cmu.edu mattias@speech.kth.se edlund@speech.kth.se

ABSTRACT

A basic requirement for participation in conversation is the ability to jointly manage interaction. Examples of interaction management include indications to acquire, re-acquire, hold, release, and acknowledge floor ownership, and these are often implemented using specialized dialog act (DA) types. In this work, we explore the prosody of one class of such DA types, known as floor mechanisms, using a methodology based on a recently proposed representation of fundamental frequency variation (FFV). Models over the representation illustrate significant differences between floor mechanisms and other dialog act types, and lead to automatic detection accuracies in equal-prior test data of up to 75%. Analysis indicates that FFV modeling offers a useful tool for the discovery of prosodic phenomena which are not explicitly labeled in the audio.

1. INTRODUCTION

A basic requirement for successful participation in conversation is the ability to jointly manage interaction. In human-human dialogue, and in multi-party settings in particular, attempts to participate frequently require negotiation; participants are faced with the dual task of deploying their contributions while at the same time limiting disruption to dialogue flow across the group. Verbal but implicit indications of intent to acquire or to retain the floor, known as *floor mechanisms*, appear crucial.

The availability of floor mechanisms as a human tactic has implications for spoken dialogue systems, which are being deployed in increasingly conversational settings. The current trend to abandon strict turn-taking and to embrace incremental and continuous processing of dialogue [1] places greater emphasis on the detection of floor mechanisms, to date only infrequently studied. Detection failures are likely to have two main consequences. First, and most obviously, failures increase the ambiguity of floor ownership. For example, mistaking floor mechanism speech for feedback may lead systems to continue speaking at a time when a human interlocutor is seeking to contribute. Second, and perhaps more importantly, the failure to react appropriately when an interlocutor has signaled his or her intent to contribute is likely to decrease the sense of mutual understanding, increasing the subsequent effort needed for system and human alike.

In this study, we investigate to what extent prosody, and intonation in particular, can be used to distinguish floor mechanism dialog acts (DAs). Prosody has previously been shown to help in DA classification [11, 13, 2]; available results are broadly useful for our purposes here, as they strive towards a general description of spontaneous speech, but reported accuracies are dominated by DA types which are frequent and long in duration. In contrast, floor mechanisms account for less than 5% of speaking time in the naturally-occurring conversations we study [12]. They also share a vocabulary with other DA types, in particular those implementing feedback, making them more difficult to distinguish lexically.

A novel aspect of the current work is its use of the fundamental frequency variation (FFV) spectrum [**?**], and standard acoustic modeling techniques as used elsewhere in speech processing, in contrast to the often arcane estimation, post-processing, and speaker normalization of pitch tracker output. We note that FFV computation has not previously been applied to audio collected outside of the anechoic chamber. Our methodology, as applied to the current task, is described in Section 3, following a description of our datasets in Section 2. Model structure is directly interpretable, as described in Section 4; furthermore, as presented in Section 5, the models can be used for automatic classification. Section 5 also contains the main contribution of this work, namely a model description of English floor mechanism prosody.

2. DATA

The data used in this work is drawn from the ICSI Meeting Corpus [7] and its associated DA annotations [12][1]. To our knowledge, it is the largest publicly available corpus of naturally-occurring unstructured multiparty conversation, consisting of longitudinal collections of meetings by several groups, and amounting to over 66 hours of meeting time. As defined in the accompanying release notes, 73 of the meetings have been divided into a TRAINSET of 51 meetings and a DEVSET and EVALSET of 11 meetings each. For our experiments, we draw training exemplars from TRAINSET and testing exemplars from DEVSET; we perform no tuning on the latter and leave EVALSET for assessment in future work.

Floor mechanisms in this data are annotated as belonging to one of three DA classes [12]; the following descriptions are taken from the guide [3] used for their annotation. **Floor grabbers** (fg) generally occur at the beginning of a speaker's turn, and tend to be louder than surrounding speech. Common lexical implementations include: "well", "and", "but", "so", "um", "uh", "I mean", "okay", and "yeah". **Floor holders** (fh) tend to occur in the middle or at the end of a speaker's turn, whereas **holds** (h) typically occur at the beginning. Prosodically, both floor holders and holds are of similar loudness with respect to surrounding speech, but of longer segment durations. Common lexical implementations of fh and h are as for fg, but also include "or", "let's see", "and what else", and "anyway". As described in [3], the three floor mechanisms share a similar vocabulary, also with backchannels and acknowledgments.

We note that there are a number of mentions of an association between flat pitch and floor holding (c.f. [4] and references therein). Also, the literature shows considerable overlap between floor mechanisms as described above and categories such as "disfluencies" and "filled pauses", which have been more thoroughly investigated (e.g. [5] and references therein). The extent of this overlap in terminology is, however, not known to us, and we have opted here to focus on the ICSI Meeting Corpus DA categories.

3. METHODOLOGY

3.1 Data Selection and Pre-processing

To facilitate subsequent discussion, we propose an intermediate temporal unit of analysis, the *talkspurt fragment* (TSF). We derive

[1] Release icsi_mrda+hs_corpus_050512.tar.gz.

this unit from two reference segmentations, both of which are available for the ICSI Meeting Corpus. The first is that of speech versus non-speech, obtained from the forced-alignment of words; we first derive from this a talkspurt (TS) segmentation, in which immediately adjacent words are merged (with no bridging of inter-word gaps). The second reference segmentation is the framing of words into DAs, which may include DA-internal pauses.

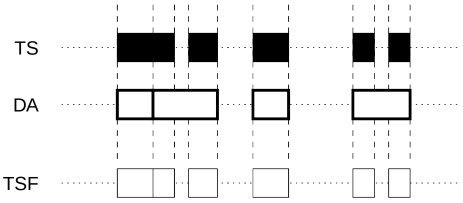


Figure 1: Construction of the talkspurt fragment (TSF), given reference talkspurt (TS) and dialog act (DA) segmentations.

An example intended to clarify our TSF definition is shown in Figure 1. A TSF is the longest unit which belongs to at most one talkspurt and at most one dialog act, and inherits the DA type of the DA to which it belongs. TSFs can be:

1. TS-initial (and optionally DA-initial), or
2. mid-TS and DA-initial; and
3. TS-terminal (and optionally DA-terminal), or
4. mid-TS and DA-terminal.

In the current work, we attempt to discriminate between floor mechanism DA types and other DA types in each of these four TSF contexts. The first two contexts are more immediately relevant to dialogue systems whose estimates of DA boundaries may be unreliable prior to lexical recognition; we include experiments involving the latter two contexts for completion. In all cases, we focus only on the 500 ms with which TSFs begin or end.

3.2 Feature Extraction

The fundamental frequency variation (FFV) representation is a 7-element characterization of within-frame variation in fundamental frequency. Its computation, which obviates the need to first estimate the fundamental frequency itself, was described in detail in [8, **?**, 10]; here, space limitations allow for only a brief account.

Following pre-emphasis ($1-0.97z^{-1}$), the signal is framed into 32 ms overlapping windows, with a frame step of 8 ms. Two frequency spectra, $\mathbf{F}_L$ and $\mathbf{F}_R$, are computed for the left and right halves of each frame, respectively, using tapered and largely disjoint windows. Each of the two spectra is then dilated in frequency, over a continuum of dilation factors, while the other spectrum is kept constant. A modified dot-product yields a measure of alignment $g(\rho)$ of their respective harmonic trains, for dilation factor ρ. Frame energy is normalized out of this representation.

We oversample the ensemble $g(\rho)$, which we refer to as the FFV spectrum, at discrete equi-spaced intervals of ρ, and then pass the resulting vector through a filterbank whose design was motivated by psychoacoustic studies [6]. An explanation of the 7 resulting filter outputs is provided in Section 4, along with an example.

3.3 Density Estimation

As mentioned in Section 3.1, the experiments we present involve discrimination for several binary partitions of the data. In each case, and for each binary class c, we estimate a hidden Markov model (HMM) $\mathscr{M}_c$ over sequences of feature vectors using maximum likelihood (ML) expectation-maximization[2] (EM), from training material belonging to that class. As in [8], we use a model of 4 fully-connected states and a single 7-dimensional Gaussian for the emission probability of each state.

3.4 Maximum Likelihood Classification

To assess the extent to which models $\mathscr{M}_c$ over the FFV representation are discriminative for unseen data, we perform automatic classification of sequences drawn from DevSet. As in [**?**], we train 10 models $\mathscr{M}_{c,i}$, $1 \leq i \leq 10$, with different random seeds for each of two binary classes, $c \in \{\alpha, \neg\alpha\}$. The classifier then assigns, to each test sequence, a class label

$$c* \;=\; \begin{cases} \alpha & \text{if } \log \frac{\prod_i P(\mathbf{x}|\mathscr{M}_{\alpha,i})}{\prod_i P(\mathbf{x}|\mathscr{M}_{\neg\alpha,i})} > \theta \\ \neg\alpha & \text{otherwise} \end{cases} \quad . \qquad (1)$$

Varying θ allows for easy construction of receiver operating characteristic curves (ROCs), which we provide in Section 5.5.

4. INTERPRETING MODEL STRUCTURE

One of our aims in this work is to illustrate the utility of hidden Markov modeling of FFV spectra as an exploratory tool. Whereas graphical depictions of finite state topologies are widespread and well understood, we describe here in some detail our depiction of the emission probabilities, an example of which is given in Figure 2.

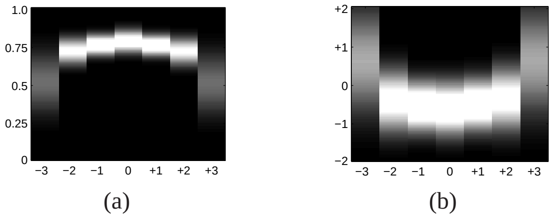


Figure 2: Depiction of single 7-dimensional Gaussian HMM emission probabilities, for (a) raw observations ($\in [0,1]$), and (b) global Z-normalized observations ($\in [\mu-2\sigma, \mu+2\sigma]$).

Figure 2(a) shows 7 vertical bands, each corresponding to the output of one filter in our filterbank. Since energy is normalized out, filterbank output values, shown on the y-axis, are bounded in $[0,1]$. Grayscale values indicate the probability of each y-axis value, per filter, given that each filter is described by one mean and one variance (cf. Section 3.3); concentrations of white designate areas which are in the vicinity of the mean, their relative widths indicate the relative magnitude of the variance, and black denotes near-zero probability.

The center band (at 0 along the x-axis) corresponds to frame-level variation of fundamental frequency in the range $[-0.1,+0.1]$ semitones per frame, or $[-1,+1]$ octaves per second, and is intended to capture those instants of speech during which the speaker may be said to be employing a flat pitch contour. Psychoacoustic studies have estimated that for a vowel 100 ms in duration, flatness is perceived when the magnitude of the observed pitch change is smaller than 16 semitones per second, or 1.33 octaves per second.

The two bands identified as -1 and $+1$ in Figure 2(a) correspond to rates of fundamental frequency change in the ranges of $[-3.4,-0.5]$ octaves per second and $[+0.5,+3.4]$ octaves per second, respectively. They describe slowly changing fundamental frequency. Quickly changing fundamental frequency is modeled by the two filters identified as -2 and $+2$ in the figure. These correspond to rates of change in the ranges of $[-5.4,-2.4]$ and $[+2.4,+5.4]$ octaves per second, respectively. Finally, the bands identified as -3 and $+3$ integrate $g(\rho)$ over the ranges of $[-2,-1]$ and $[+1,+2]$ octaves *per 0.008 ms*. They account in part for the anticipated correlations between $\mathbf{F}_L$ and $\mathbf{F}_R$ due to frequency halving and doubling during voiced speech. During unvoiced speech, all filter outputs exhibit much larger variance and are less correlated than during voiced speech.

[2] Kevin Murphy's implementation in Matlab™, available at `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html` was used for all HMM operations (downloaded on Feb 9 2009).

As described, Figure 2(a) shows an example model over the raw output of the filterbank. For highlighting differences between competing HMMs, we instead use models over Z-normalized values, shown in Figure 2(b). A global Z-transform is computed for each filter's response across all TRAINSET frames used in a particular experiment.

5. EXPERIMENTS AND ANALYSIS

5.1 Talkspurt initiation (TSI)

In a first experiment, we group together all floor mechanism DA types: floor grabbers (fg), floor holders (fh), and holds (h), and attempt to determine whether they differ from other DA types, ¬{fg+fh+h}, immediately when they start. We do this by training ML models of the first 500 ms of TS-initial TSFs for both classes. Both the {fg+fh+h} and the ¬{fg+fh+h} models are trained using 7500 exemplars, drawn randomly from the total number available; this number is limited by the number of floor mechanism instances in TRAINSET. We use the trained models to classify a set of 1500 TSFs drawn from DEVSET for both classes.

The results of this experiment, for which random guessing would lead to an accuracy of 50% and a ROC discrimination of 50%, are shown in the first line of Table 1. It can be seen that accuracy is significantly above random guessing for TRAINSET TSFs, and that the results generalize well to unseen data (DEVSET).

We also design a similar experiment attempting to differentiate between {fh+h}, which we expect to be internally similar (and dissimilar from floor grabbers), and ¬{fh+h}. We use 7000 TSFs for training each of the two models, and 1000 TSFs for testing. As the second line of the first panel in Table 1 shows, this less diverse set of DA phenomena is easier to differentiate from the rest, and generalization to unseen data remains high.

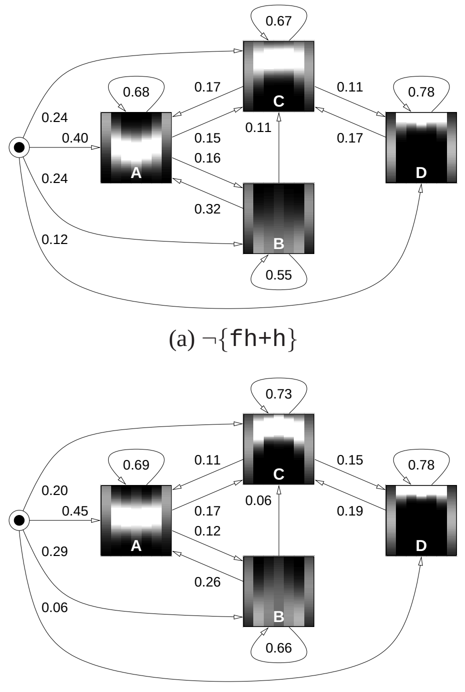
In a third experiment, we attempt to differentiate between h and ¬h; because instances of h are rare, only 500 and 180 TSFs were available for training and testing, respectively. As can be seen in the third line of the first panel of Table 1, h instances are significantly more differentiable from ¬h instances than are {fh+h} instances from ¬{fh+h} instances. We also note that generalization is slightly poorer than in the previous two experiments, a fact we attribute to the smaller number of training exemplars.

Context	DA types to detect	Acc. (%) at $\theta = 0$		ROC area
		TRAIN	DEV	DEV
TSI (§5.1)	{fg,fh,h}	62.7	63.9	70.2
	{fh,h}	65.1	65.3	72.2
	h	74.6	72.7	82.0
TST (§5.2)	{fg,fh,h}	59.2	58.3	63.5
	{fh,h}	63.5	64.1	70.6
	h	72.4	68.3	75.6
DAI (§5.3)	fh	62.4	63.6	68.4
DAT (§5.4)	fh	66.7	67.4	75.0

Table 1: Binary classification accuracies (using Equation 1) for TRAINSET and DEVSET, as well as discrimination (in %, with maximum value of 100%) for several TSF contexts.

In Figure 3, we show the transition and Z-normalized emission probabilities of models trained on the {fh+h} and ¬{fh+h} TRAINSET data. As can be seen, the transition probabilities for both state networks are very similar, as are the emission probabilities for states labeled "B" and "D". However, for the most likely first state, labeled "A", the emission probability models differ. For ¬{fh+h}, in diagram (a), the response of the central filter, and of the ±1 filters, is lower than its global average, while that of the ±2 filters is slightly above their global average. In (b), for {fh+h}, the responses of all five filters are slightly above their global averages. This indicates a lower probability of the use of flat or slowly varying pitch upon entering ¬{fh+h} talkspurts, relative to that employed when entering {fh+h} talkspurts.

Also, we note that the central filter response in the state labeled "C" in both diagrams is higher for {fh+h}, in (b), than it is for ¬{fh+h}, in (a). This indicates a higher likelihood of flat pitch use for {fh+h} talkspurts, in the state most likely the next to be visited following egress from state "A".



(a) ¬{fh+h}

(b) {fh+h}

Figure 3: Models inferred for TS-initial talkspurt fragments; Z-normalized emission probabilities shown.

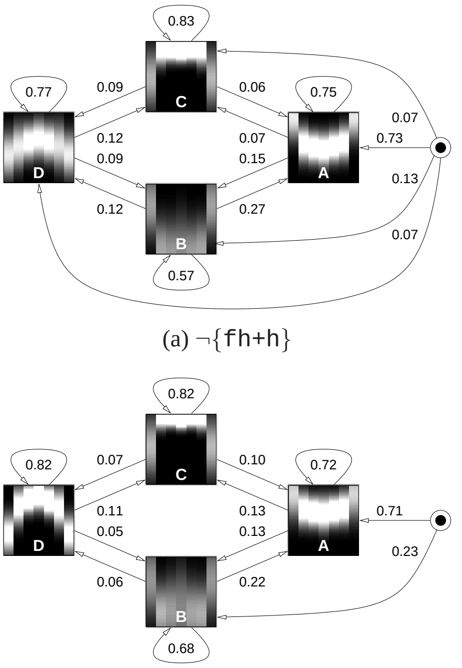
5.2 Talkspurt termination (TST)

Having treated talkspurt initiation in Section 5.1, we now turn to talkspurt termination, and revisit the exact same distinctions among DA type classes. In this case, models are trained on the *last* 500 ms of each talkspurt-terminal TSF, reversed in time such that the last frame appears first in the sequence. Time reversal is performed only to facilitate analysis, yielding in our framework a single egress state out of talkspurt fragments, rather than a single entry state. The number of training and testing instances is 7000 and 1500, respectively, for the {fg+fh+h} versus ¬{fg+fh+h} task; 5000 and 1000, respectively, for the {fh+h} versus ¬{fh+h} task; and 320 and 120, respectively, for the h versus ¬h task.

The results for all three experiments are shown in the second panel of Table 1. As can be seen, at talkspurt-terminal locations automatic classification yields accuracies which are significantly above random guessing (of 50%), which generalize to unseen data, and which follow approximately the same trend as at talkspurt-initial locations as the class of interest is narrowed progressively towards purely h speech. However, all accuracies are slightly lower than they are for talkspurt-initial classification, suggesting that it is more difficult to differentiate floor mechanisms using variation of fundamental frequency as they are ending.

Figure 4 shows model structure for both {fh+h} and ¬{fh+h} classes, in diagrams (a) and (b), respectively, trained in the same way as in the preceding section. We again note that the transition probabilities in both diagrams (a) and (b) are quite similar. Although the most visible difference is that between the emission probabilities of states labeled "D", these states are not frequently visited; we focus instead on the remaining three states.

As can be seen, the relative response of all filters for these three states is visually quite similar between diagrams (a) and (b). However, closer inspection reveals that the Gaussian means are all higher in (b) than in (a). This result indicates that $g(\rho)$ attains higher values while {fh+h} speech is terminating than while ¬{fh+h} is terminating, regardless of F0 slope. Our explanation for this is as follows. In computing $g(\rho)$, we do not normalize out the spectral envelope of $\mathbf{F}_L$ or of $\mathbf{F}_R$; as a result, $g(\rho)$ can be expected to be higher when the envelopes are more similar. We believe that in instances of identical filterbank response *shape*, a higher *offset* for all filters indicates more slowly varying spectral envelope, i.e. a slower rate of speaking. This observation is corroborated by the MRDA annotation manual [3], but was unexpected from the FFV representation. Evidently, the effect is strong enough to form a basis for discriminating among these two classes of DAs at talkspurt ends.



(a) ¬{fh+h}

(b) {fh+h}

Figure 4: Models inferred for TS-terminal talkspurt fragments; *Z*-normalized emission probabilities shown.

5.3 Mid-talkspurt DA initiation (DAI)

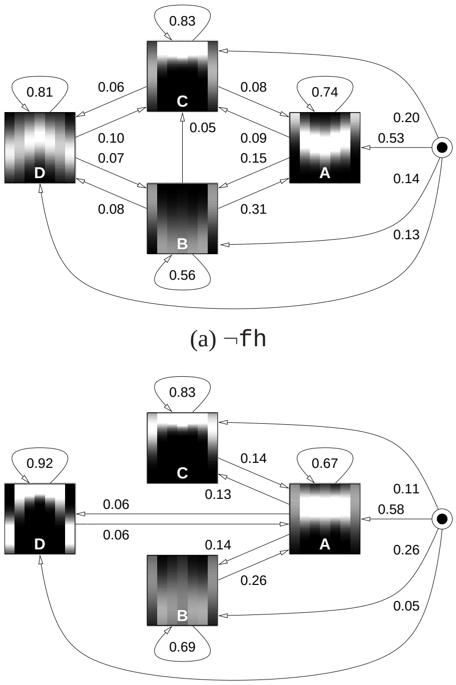
We now turn to two complementary problems, in this section and the next, asking whether floor mechanisms which begin and end in the middle of an ongoing talkspurt can be distinguished from other types of DAs. This scenario is less directly applicable to conversational dialogue systems, whose estimates of DA boundaries at runtime are likely to be less dependable than those of talkspurt boundaries. In this section, we focus on 500 ms of mid-talkspurt DA-initial speech; only the fh DA type occurs frequently enough in this context to allow for modeling. We use 1200 TSFs for training both class models, and 240 TSFs from each class for testing.

The results, given in the third panel of Table 1, indicate performance which is on par (but slightly lower) with that for TS-initial contexts for all floor mechanisms ({fg+fh+h} in panel 1). Due to space constraints, we do not show the model structure for this case. However, we note that transition probabilities for both networks are similar, and, as in the previous section, the only visually apparent difference is that the offsets for filter responses of the fh model are higher than for those of the ¬fh model. Given our current understanding, we believe that the models are differentiating among mid-TS fh and ¬fh initiation based largely on speaking rate.

5.4 Mid-talkspurt DA termination (DAT)

Finally, we explore the same distinction as in Section 5.3, but for DAs which end in mid-TS. As for that task, only fh DAs occur sufficiently frequently for modeling; we use 750 TSFs for training both the fh and the ¬fh models, and 180 testing TSFs from each class drawn from DevSet. The results, shown in the fourth panel of Table 1, indicate that in this location fh closure is easier to detect than its onset; furthermore, performance is also better than for the {fg+fh+h} versus ¬{fg+fh+h} and {fh+h} versus ¬{fh+h} distinctions in the second panel of the table.

The models learned for this task are shown in Figure 5. Although there are several differences between the ¬fh model and the fh model to note, we focus in particular on the most likely state to terminate the TSFs in question, identified as "A" in both diagrams. This state terminates both types of TSFs over 50% of the time. As can be seen, the response of the center filter of state "A" in the ¬fh model is approximately what it is on average, while the response of the slowly varying fundamental frequency filters is higher than it is on average. This is even more the case for the quickly varying fundamental frequency filters. In contrast, for the fh model, the response of all filters is above average in the state labeled "A". This difference between the two models suggests that flat pitch is used to terminate mid-TS fh TSFs more frequently, in relative terms, than it is used to terminate mid-TS ¬fh TSFs.



(a) ¬fh

(b) fh

Figure 5: Models inferred for mid-TS DA-terminal ¬fh (a) and fh (b) TSFs; *Z*-normalized emission probabilities shown.

We note in passing here that the state labeled "D" in both diagrams in Figure 5 also exhibits quite different emission probabilities; however, this state is only rarely visited, in both cases.

5.5 Task Comparison

The accuracies presented in Table 1 characterize performance when the log-likelihood-ratio threshold θ is zero; to describe performance over the full range of θ, corresponding to different weights associated with precision and recall, we present receiver operating characteristic curves for the four pairs of models shown in Figures 3 through 5, in Figure 6. Over most of the range of possible true positive rates, the four detection tasks appear to be increasingly more difficult in the order: mid-TS DA-terminal fh TSFs, TS-initial fh+h, TS-terminal fh+h, and mid-TS DA-initial fh. However, at high true positive rates, the two tasks which appear to rely more on speaking rate than on variation in fundamental frequency outperform the other two. This effect is currently under investigation.

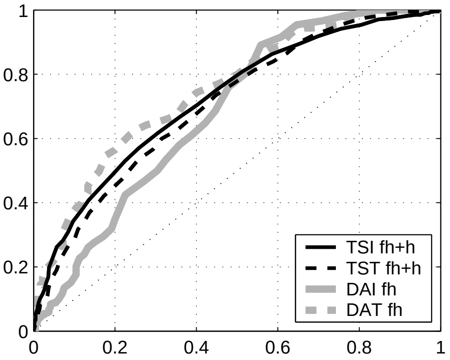


Figure 6: Receiver operating characteristic curves for detection using Equation 1 and the models of Sections 5.1, 5.2, 5.3, and 5.4. The $x-$ and $y-$axes represent false and true positive rates, respectively.

6. DISCUSSION

The experiments presented indicate that floor holders fh and holds h (which together make up the majority of floor mechanism production by time) differ from other DAs in the rate of F0 variation and/or the rate of speech, at both their onsets and ends, regardless of whether they begin or terminate at talkspurt boundaries.

The models we infer for the classes described are single multivariate Gaussian distributions with diagonal covariances. When change in inter-harmonic spacing is slow, the models appear to confound the effects of change in harmonic structure and change in spectral envelope. We believe that $g(\rho)$ is dominated by changes in harmonic structure, both from an underlying mathematical perspective as well as from the fact that we achieve higher classification accuracies where inferred model pairs differ in filterbank response shape rather than offset. Shape normalization by subtraction of the offset, and separate modeling of the offset, may in the future shed light on which effect is more pronounced, for a given context. A more optimal strategy may involve inverse filtering $\mathbf{F}_L$ and $\mathbf{F}_R$ prior to the computation of $g(\rho)$.

We note for completion that the floor mechanism TSFs used in our experiments are on average shorter than other DA type TSFs. When a particular TSF is shorter than 500 ms, we truncate the audio, leading to shorter sequences of feature vectors; not doing so would encourage models to learn silence after the end of TS-initial TSFs and before the start of TS-terminal TSFs. Separate experiments using only a duration threshold have shown that for h detection, for example, accuracy is barely above 50%. This indicates that audio truncation by itself is not responsible for the classification accuracies observed using FFV models.

7. CONCLUSIONS

We have demonstrated the use of the fundamental frequency variation (FFV) representation in exploratory analysis of a rare but important phenomenon in spontaneous speech. The work presented is the first application of FFV modeling to the inference of dialogue act type in multiparty conversation, on non-anechoic-chamber, close-talk-microphone recordings which exhibit a significant amount of crosstalk. To a first approximation, our experiments suggest that, at talkspurt-initial locations, floor holders and holds differ from other DA types in their more frequent use of flat pitch. Talkspurts belonging to these two classes appear to terminate with the use of slower speech. Conversely, in mid-talkspurt locales, floor holders appear to begin with slower speech than do other DA types, but end with relatively more frequent use of flat pitch. These findings corroborate, and elaborate on, those in the literature regarding floor holding mechanisms, and support our previous finding that flat pitch is strongly predictive of locations in which interlocutors in two-party dialogue do not begin to speak. Overall, the proposed methodology has been shown to be suitable for the discovery and analysis of prosodic phenomena which, as in this work, are not explicitly labeled. We expect that it will service future automatic discrimination efforts as well as improved human understanding.

8. ACKNOWLEDGMENTS

We would like to thank Liz Shriberg for access to the ICSI MRDA Corpus. This work was funded in part by the Swedish Research Council project #2006-2172 *What makes speech special?*.

REFERENCES

[1] J. Allen, G. Ferguson, and A. Stent, “An architecture for more realistic conversational systems, ” in *Proc. IUI*, Santa Fe NM, USA, January 14 – 17, 2001. pp. 1–8.

[2] J. Ang, Y. Liu, and E. Shriberg, “Automatic dialog act segmentation and classification in multiparty meetings, ” in *Proc. ICASSP*, 2005. pp. 1061-1064.

[3] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, “Meeting Recorder Project: Dialog act labeling guide, ” *ICSI Technical Report TR-04-002*, 2004.

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

[5] R. Eklund, “Disfluencies in Swedish human-human and human-machine travel booking dialogues. ” in *Linköping Studies in Science and Technology*, Dissertation No. 882, Linköping, Sweden, 2004.

[6] J. ’t Hart, R. Collier, and A. Cohen, *A perceptual study of intonation: An experimental-phonetic approach to speech melody*, Cambridge University Press, 1990.

[7] A. Janin et al, “The ICSI Meeting Corpus, ” in *Proc. ICASSP*, 2003. pp. 364–367.

[8] 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, March 30 – April 04, 2008. pp. 5041–5044.

[9] K. Laskowski, J. Edlund, and M. Heldner, “Learning prosodic sequences using the fundamental frequency variation spectrum, ” in *Proc. Speech Prosody*, Campinas, Brazil, May 06 – 09, 2008.

[10] K. Laskowski, M. Wölfel, M. Heldner, and J. Edlund, “Computing the fundamental frequency variation spectrum in conversational spoken dialogue systems, ” in *Proc. ACOUSTICS*, Paris, France, June 29 – July 04, 2008. pp. 3305–3310.

[11] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, et al, “Can prosody aid in the automatic classification of dialog acts in conversational speech?” in *Language and Speech 41(3–4)*, 1998. pp. 443–492.

[12] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, H. Carvey, “The ICSI MRDA Corpus” in *SIGdial*, 2004. pp. 97–100.

[13] A. Venkatamaran, L. Ferrer, A. Stolcke, and E. Shriberg, “Training a prosody-based dialog act tagger from unlabeled data” in *Proc. ICASSP*, 2003. pp. 272–275.