Computing the fundamental frequency variation spectrum in conversational spoken dialogue systems

K. Laskowski[a], M. Wölfel[b], M. Heldner[c] and J. Edlund[c]

[a]Carnegie Mellon University, 407 South Craig Street, Pittsburgh, PA 15213, USA
[b]Universität Karlsruhe, am Fasanengarten 5, 76131 Karlsruhe, Germany
[c]KTH, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden
kornel@cs.cmu.edu

Continuous modeling of intonation in natural speech has long been hampered by a focus on modeling fundamental frequency, of which several normative aspects are particularly problematic. The latter include, among others, the fact that pitch is undefined in unvoiced segments, that its absolute magnitude is speaker-specific, and that its robust estimation and modeling, at a particular point in time, rely on a patchwork of long-time stability heuristics. In the present work, we continue our analysis of the fundamental frequency variation (FFV) spectrum, a recently proposed instantaneous, continuous, vector-valued representation of pitch variation, which is obtained by comparing the harmonic structure of the frequency magnitude spectra of the left and right half of an analysis frame. We analyze the sensitivity of a task-specific error rate in a conversational spoken dialogue system to the specific definition of the left and right halves of a frame, resulting in operational recommendations regarding the framing policy and window shape.

1   Introduction

Variation in prosody, including loudness, pitch, and tempo, is an important aspect of human interaction and dialogue, and spoken dialogue systems aiming at mimicking human vocal behavior must model such variation. Examples of tasks in which the use of prosodic models has been explored for dialogue systems include identification of places to use back-channel feedback [10], classification of rhetorical relations [7], interpretation of discourse markers (e.g. [5]), dialogue act tagging [9], and identification of speech repairs [1, 6].

The continuous modeling of intonation, or pitch variation, in natural speech has long been hampered by a focus on modeling fundamental frequency (F0), of which several normative aspects are particularly problematic. The latter include, among others, the fact that F0 is undefined in unvoiced segments, that its absolute magnitude is speaker-specific, and that its robust estimation and modeling, at a particular point in time, rely on a patchwork of long-time stability heuristics. In the current work, we continue our analysis of a direct measure of *variation* in F0, the fundamental frequency variation (FFV) spectrum [3, 4], which does not require F0 estimation. In particular, the FFV spectrum is an instantaneous, continuous, vector-valued representation of F0 variation, which incurs few if any of of the above limitations.

We first provide a description of the estimation and modeling of the FFV spectrum, in the context of an end-to-end dialogue system component. Our account is particularly focused on details which have received little attention in our previous work, namely the effect of the windowing policy on component performance. We then present several experiments which justify our initial design considerations, but which also indicate several avenues for improvement in future work. It is our intention that our descriptions in [3, 4] and the detail provided herein be sufficient for others to successfully reproduce our end-to-end component.

2   Speaker-Change Prediction in Spoken Dialogue Systems

The task which we explore in the current work is prediction. Given a pause of 0.3 seconds or longer following a talkspurt [8] from one participant in spontaneous human-human dialogue, we train a classifier to predict which of the two participants will speak next. We de-fine the other participant speaking next as a *speaker-change* (SC), and the situation in which the same participant speaks next as *not-a-speaker-change* ($\neg$SC). Binary classification is based on the last 500ms of acoustic signal prior to the end of the talkspurt; we refer to this as the end-of-talkspurt (EOT) condition. We also explore the contrastive, diagnostic end-of-voicing (EOV) condition, in which we use 500ms of acoustic signal prior to the end of the last voicing interval in the talkspurt in question. Details regarding this setup can be found in [3].

Data used in our experiments, as in our earlier work, has been drawn from the Swedish Map Task Corpus [2]. It consists of a DevSet with 480 EOTs, of which 222 are SCs. This data is used for model training, as well as round-robin development testing. The EvalSet contains 317 EOTs of which 149 are SCs. Importantly, the two data sets are disjoint in speakers.

3   System Operation

The proposed SC/$\neg$SC classifier runs potentially in parallel with other audio processing applications. The current implementation expects, from Step 0 of Figure 1, a stream of audio sampled at 16 kHz. The audio is then pre-emphasized using a standard FIR filter, shown as Step 1.

At Step 2, the audio is framed at a rate of 125 Hz (a frame increment of $t_{fra} = 0.008$ seconds), using a frame size of 0.032 seconds. Spectral estimation is performed for the left and right halves of the frame, using a pair of windows; this step is described in detail in Section 4. The two spectra, F_L and F_R, are assumed to represent the spectal content of the signal at the temporal locations $t_0 - t_{sep}/2$ and $t_0 + t_{sep}/2$, respectively, where t_0 is the center of the frame, and t_{sep} is the separation between the maxima of the two windows.

The fundamental frequency variation spectrum is computed at Step 3, using F_L and F_R; the resulting spectrum, of 256 elements, is considered as characterizing variation over an interval of t_{sep} seconds. It is then compressed using a filterbank (shown in Figure 3 in [4]), in Step 4, leading to a 7-scalar representation per frame. This representation is centered and rotated in Step 5, using a whitening transform learned from training data.

In parallel with Steps 1 (or 2) through 5, a speech activity detector (SAD) is applied to the audio from Step 0 (or 1). When the SAD subsystem signals that a talkspurt has ended, Step 6 in Figure 1 computes the likelihood of the last 500 ms of frames from Step 5 (62
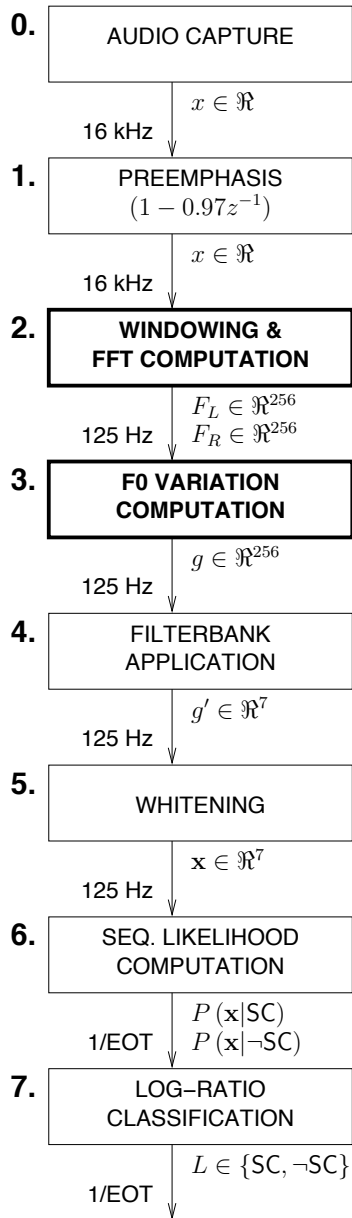
4 Windowing and FFT Computation

Computation of a left and a right spectrum, per analysis frame, is performed using two window functions, a left and a right counterpart. The shapes of the two windows are mirror images of each other, when reflected about the center of the frame t_0. They are shown in Figure 2; for brevity, we describe only the left-side window.
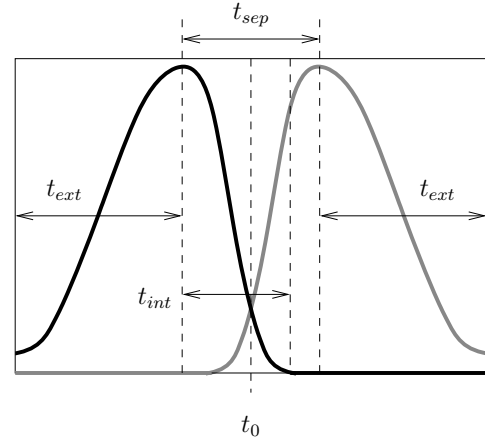

Figure 2: Nominal shape and location of the h_L and h_R windows within a single analysis frame of width t_{wid}, centered on t_0; symbols as in the text.

As the peaks of the two windows are t_{sep} apart, the peak of the left window can be found at $t_0 - t_{sep}/2$. The exterior edge of the window (the left edge for the left window, and vice versa for the right window), is non-zero for t_{ext} seconds from the peak; the interior edge is non-zero for t_{int} seconds from the peak. This means that the the width of the analysis frame, from the leftmost non-zero value of the left window to the rightmost non-zero value of the right window, is $t_w = t_{ext} + t_{sep} + t_{ext}$. There may be partial temporal overlap between the left and the right window, given by

$$
\begin{aligned}
t_{ol} &= \left(-\frac{t_{sep}}{2} + t_{int}\right) - \left(+\frac{t_{sep}}{2} - t_{int}\right) \\
&= 2t_{int} - t_{sep}
\end{aligned}
\tag{1}
$$

Finally, we note that the shapes of the exterior and interior edges of both windows are borrowed from the Hamming and Hann windows, respectively. The two windows are therefore completely specified by the parameters t_{int}, t_{ext}, and t_{sep}; consecutive frames are computed t_{fra} seconds apart.

In our previous work [3, 4], we had collapsed these 4 parameters to 2 parameters, T_0 and T_1. The latter can be expressed in terms of the former,

$$
T_0 = \frac{t_{sep}}{2} = \frac{t_{int}}{2} = \frac{t_{fra}}{2} \tag{2}
$$

$$
T_1 = t_{ext} \tag{3}
$$

The main focus of the current work is to explore the impact of t_{int}, t_{ext}, and t_{sep} on SC/¬SC classification. t_{fra} will be kept fixed at 0.008 seconds. Additionally, modifications of the remaining three parameters will be

Figure 1: SC/¬SC prediction system architecture. Data rate and format depicted between consecutive components. Steps 2 and 3, the focus of the current work, are shown in bold. The final date rate of 1/EOT indicates that L predictions are available only as often as end-of-talkspurts (EOTs) are.

of them at a frame rate of 125 Hz), given competing sets of 10 hidden Markov models for each of SC and ¬SC. Using a log-likelihood-ratio classifier, the system outputs a hypothesis $L_t \in \{\text{SC}, \neg\text{SC}\}$ in Step 7, for a talkspurt ending at time t. The hypothesis is then used to inform immediate dialogue system strategy.

Details regaring the operation of Steps 4 through 7 can be found in [4]. The focus of the current work is Step 2, responsible for spectral estimation. Additionally, to avoid redefining the filterbank of Step 4 each time the parameter t_{sep} is modified, we offer a more detailed description of Step 3 than presented in our earlier work.

subject to the constraint that the overall width of the analysis frame, t_w, has constant value of 0.032 s. These two critera ensure that in each experiment, the ultimate 500 ms of every talkspurt are represented by 62 consecutive FFV spectra, each obtained from two 512-point FFTs.

5 Computation of the Fundamental Frequency Variation Spectrum

The FFV spectrum g was derived in [3], using geometric arguments. It involved an independent variable τ, and the two continuous spectra F_L and F_R:

$$
g(\tau) = \tag{4}
$$
$$
\begin{cases}
\int_{-f_s/2}^{+f_s/2} F_L\left(\left(\frac{-\tau-T_0}{-\tau+T_0}\right)f\right) F_R^*(f)\, df & \tau < -T_0 \\
\int_{-f_s/2}^{+f_s/2} F_L(f) F_R^*\left(\left(\frac{+\tau-T_0}{+\tau+T_0}\right)f\right) df & \tau > +T_0
\end{cases}
$$

Although instructive during the derivation, τ represents an inconveniently discontinous domain for g^τ, which is undefined over the interval $[-T_{sep}/2, +T_{sep}/2]$. A support for g which is continuous over $(-\infty, +\infty)$ — and, optimally, analytic — is preferable. We therefore define the conformal mapping

$$
\alpha : \tau \mapsto \rho =
\begin{cases}
-\log_2\left(\frac{-\tau-t_{sep}/2}{-\tau+t_{sep}/2}\right), & \tau < -t_0 \\
+\log_2\left(\frac{+\tau-t_{sep}/2}{+\tau+t_{sep}/2}\right), & \tau > +t_0
\end{cases}. \tag{5}
$$

Eq (5) offers the additional advantage that while $2^{\pm\rho}$ represents change in *linear frequency* per separation lag t_{sep}, ρ itself represents the same change in *octaves* per separation lag. We illustrate the mapping at sample values of τ in Table 1.

Table 1: Sample (τ, ρ) pairs in conformal mapping $\alpha : \tau \mapsto \rho$, which allows for reexpression of the derived g^τ as g^ρ, a function of variation in octaves per second.

τ	$+\frac{5}{6}t_{sep}$	$+3t_{sep}$	$+\infty$	$-\infty$	$-3t_{sep}$	$-\frac{5}{6}t_{sep}$
ρ	-2	-1	$\to 0 \leftarrow$	$+1$	$+2$

Reexpression of Eq(4) in terms of ρ then yields

$$
g^\rho(\rho) =
\begin{cases}
\int_{-f_s/2}^{+f_s/2} F_L(f) F_R^*(2^{+\rho}f)\, df, & \rho < 0 \\
\int_{-f_s/2}^{+f_s/2} F_L(2^{-\rho}f) F_R^*(f)\, df, & \rho \geq 0
\end{cases}, \tag{6}
$$

where the ρ superscript in g^ρ indicates redefinition in the mapped domain. In practice, we compute Eq (6) using magnitude rather than complex spectra,

$$
g^\rho(\rho) =
\begin{cases}
\int_{-f_s/2}^{+f_s/2} |F_L(f)| |F_R^*(2^{+\rho}f)|\, df, & \rho < 0 \\
\int_{-f_s/2}^{+f_s/2} |F_L(2^{-\rho}f)| |F_R^*(f)|\, df, & \rho \geq 0
\end{cases}. \tag{7}
$$

In the proposed implementation, $F_L[k]$ and $F_R[k]$ are discrete, complex, N=512-point FFT representations, defined for $k \in Z$ restricted to $[-N/2, +N/2]$. For the current purposes, we may substitute integration

in Eq (7) with summation over those values of f where $f = f_s k/N$, that is

$$
|F_L(f_s k/N)| = |F_L[k]|, \tag{8}
$$
$$
|F_R^*(f_s k/N)| = |F_R^*[k]|. \tag{9}
$$

However, the discrete transforms $F_L[k]$ and $F_R[k]$ are in general not defined at the corresponding dilated frequencies $2^{\pm\rho}f$. To address this problem, we resort to linear interpolation using the coefficients

$$
\beta_-^{\rho,k} = \left|\lceil 2^{-\rho}k\rceil - 2^{-\rho}k\right|, \tag{10}
$$
$$
\beta_+^{\rho,k} = \left|\lceil 2^{+\rho}k\rceil - 2^{+\rho}k\right|. \tag{11}
$$

With the help of Eqs (10&11), we then define

$$
\left|\tilde{F}_R^*(2^{+\rho}k)\right| \equiv \left|F_R^*(2^{+\rho}f_s k/N)\right| \tag{12}
$$
$$
= \beta_+^{\rho,k}\left|F_R^*\left[\lfloor 2^{+\rho}k\rfloor\right]\right| + (1-\beta_+^{\rho,k})\left|F_R^*\left[\lceil 2^{+\rho}k\rceil\right]\right|,
$$
$$
\left|\tilde{F}_L(2^{-\rho}k)\right| \equiv \left|F_L(2^{-\rho}f_s k/N)\right| \tag{13}
$$
$$
= \beta_-^{\rho,k}\left|F_L\left[\lfloor 2^{-\rho}k\rfloor\right]\right| + (1-\beta_-^{\rho,k})\left|F_L\left[\lceil 2^{-\rho}k\rceil\right]\right|.
$$

Eqs (8,9,12&13) allow us to reformulate Eq (7) as

$$
g^\rho(\rho) =
\begin{cases}
\sum |F_L[k]| \left|\tilde{F}_R^*(2^{+\rho}k)\right|, & \rho < 0 \\
\sum \left|\tilde{F}_L^*(2^{-\rho}k)\right| |F_R^*[k]|, & \rho \geq 0
\end{cases}. \tag{14}
$$

For clarity, we have elided the summation limits, $k \in [-N/2, +N/2]$. As we desire an *energy-independent* representation of fundamental frequency variation, we apply a standard spherical normalization to obtain

$$
g_N^\rho(\rho) =
\begin{cases}
\frac{\sum |F_L[k]| \left|\tilde{F}_R^*(2^{+\rho}k)\right|}{\sqrt{\sum |F_L[k]|^2 \sum \left|\tilde{F}_R^*(2^{+\rho}k)\right|^2}}, & \rho < 0 \\
\frac{\sum \left|\tilde{F}_L^*(2^{-\rho}k)\right| |F_R^*[k]|}{\sqrt{\sum \left|\tilde{F}_L^*(2^{-\rho}k)\right|^2 \sum \left|F_R^*[k]\right|^2}}, & \rho \geq 0
\end{cases}. \tag{15}
$$

Finally, the question remains at which discrete values of ρ to evaluate g_N^ρ. In previous work [3, 4], we computed the FFV spectrum at

$$
\rho = \frac{4r}{N} \tag{16}
$$

for $r \in Z$, restricted to $r \in [-N/2, +N/2]$. Since this domain is the same as for k, the operation yields a real N=256-point vector. The leftmost and rightmost values represent the magnitude of g_N^ρ for a FFV delta of ± 2 octaves per t_{sep} seconds.

We modify the sampling policy in the current work. As mentioned in Section 3, in subsequent dimensionality reduction of g_N^ρ, the filterbank filters of Step 4 in Figure 1 are defined in terms of discrete values of FFV delta in octaves per second. Ordinarily, the filter specifications would need to be modified when t_{sep} changes. To avoid this, that is to keep the filterbank structurally identical across the experiments conducted here, we instead evaluate g_N^ρ at the values

$$
\rho = \frac{4r}{N} \cdot \frac{t_{sep}}{t_{sep}^{ref}} \tag{17}
$$

where $t_{sep}^{ref} = 0.008$ s is the value with which the fixed filterbank has been defined in our previous work [3, 4].

6 Experiments

We first analyze the sensitivity of classification accuracy to the amount of overlap t_{ol} between the two windows h_L and h_R. t_{ext} and t_{sep} are kept fixed at values 0.012 s and 0.008 s, respectively; only t_{int} is modified.

The results are shown in Table 2, using 4 discrimination measures. Column heading "prod" refers to the discrimination achieved by using the product of all 10 HMMs for both SC and ¬SC as the likelihood of all pre-EOT or pre-EOV 500 ms sequences. In contrast, "mean" shows the average discrimination, over all 100 pairs consisting of one of the 10 SC HMMs and one of the 10 ¬SC models; "min" and "max" are the smallest and largest discrimination areas (whose maximum value in % is 50.0), over the same 100 pairs.

Table 2: SC/¬SC discrimination (area above the ROC diagonal; maximum value is 50.0) for the DEVSET and EVALSET, in %, as a function of the interior extent of windows h_L and h_R, for both the EOV and EOT condtion. The corresponding window overlap t_{ol} is also shown. In all conditions, $t_{ext} = 0.012$ s and $t_{sep} = 0.008$ s. "prod", "max", "min", and "mean" as described in the text.

t_{int} (s)	t_{ol} (s)	prod	max	min	mean
DEVSET, *EOV condition*					
0.006	0.004	6.2	9.7	-1.6	4.7
0.008	0.008	**10.9**	**11.5**	**4.3**	**7.9**
0.010	0.012	7.3	8.8	3.5	6.1
0.012	0.016	8.1	9.3	1.2	5.9
0.014	0.020	5.9	9.0	0.4	5.1
DEVSET, *EOT condition*					
0.006	0.004	**7.4**	10.0	-0.2	**6.6**
0.008	0.008	5.8	**10.2**	**1.9**	5.7
0.010	0.012	3.7	9.0	-0.9	3.3
0.012	0.016	3.9	7.9	0.5	3.4
0.014	0.020	5.8	8.6	1.5	4.7
EVALSET, *EOV condition*					
0.006	0.004	**18.3**	14.9	5.1	11.5
0.008	0.008	18.0	**18.9**	**9.0**	**13.9**
0.010	0.012	15.2	17.2	5.8	12.8
0.012	0.016	14.8	15.8	7.8	11.7
0.014	0.020	14.6	16.1	6.3	11.2
EVALSET, *EOT condition*					
0.006	0.004	11.9	20.1	8.4	12.2
0.008	0.008	**19.4**	**21.5**	10.2	**17.0**
0.010	0.012	18.7	**21.5**	**11.5**	16.2
0.012	0.016	17.6	20.1	7.0	13.1
0.014	0.020	13.4	17.8	6.2	10.4

Table 2 shows that the best performance, using all measures, is generally to be found for $t_{int} \in [0.006, 0.010]$ s, and not at $t_{int} = 0.012$ s, for which the left and right non-zero support of both h_L and h_R are identical. We note that although $t_{int} = 0.006$ s occasionally outperforms $t_{int} = 0.008$ s on the "prod" measure, it does not do so on "min", suggesting that $t_{int} = 0.006$ s can lead to less discriminative pairs of HMMs for the two classes.

Variation of behavior for DEVSET in both conditions, as a function of t_{int}, is similar for EVALSET; the relative increase in "prod" discrimination from $t_{int} = 0.012$ s (symmetric support windows) to $t_{int} = 0.008$ s is 35%, 49%, 22%, and 10%, for DEVSET/EOV, DEVSET/EOT, EVALSET/EOV, and EVALSET/EOT, respectively.

Second, we analyze the sensitivity of classification accuracy to the separation between the two windows, t_{sep}. t_{int} is modified accordingly to ensure that $t_{ol} = 0.008$ s across this suite of experiments. Similarly, t_{ext} is modified such so as to keep $t_w = 2t_{ext} + t_{sep}$ constant, and equal to 0.032 s. The results are shown in Table 3.

Table 3: SC/¬SC discrimination (area above the ROC diagonal; maximum value is 50.0) for the DEVSET and EVALSET, in %, as a function of the window peak separation t_{sep}, for both the EOV and EOT condtion. The corresponding internal and external support of h_L and h_R is also shown. "prod", "max", "min", and "mean" as described in the text.

t_{sep} (s)	t_{int} (s)	t_{ext} (s)	prod	max	min	mean
DEVSET, *EOV condition*						
0.008	0.008	0.012	**10.9**	**11.5**	4.3	**7.9**
0.010	0.009	0.011	6.2	9.3	0.2	5.7
0.012	0.010	0.010	6.9	9.6	2.8	6.1
0.014	0.011	0.009	7.4	9.0	**6.9**	7.6
DEVSET, *EOT condition*						
0.008	0.008	0.012	5.8	**10.2**	1.9	5.7
0.010	0.009	0.011	**7.4**	7.9	1.5	5.6
0.012	0.010	0.010	7.2	8.7	**3.7**	**6.6**
0.014	0.011	0.009	4.4	5.9	3.4	4.8
EVALSET, *EOV condition*						
0.008	0.008	0.012	18.0	18.9	**9.0**	13.9
0.010	0.009	0.011	**19.4**	**20.0**	8.1	15.1
0.012	0.010	0.010	19.3	19.5	6.0	**17.5**
0.014	0.011	0.009	17.6	18.7	3.3	16.9
EVALSET, *EOT condition*						
0.008	0.008	0.012	19.4	**21.5**	10.2	17.0
0.010	0.009	0.011	19.0	21.1	10.5	16.9
0.012	0.010	0.010	**20.7**	21.0	13.1	20.0
0.014	0.011	0.009	**20.7**	21.2	**20.4**	**20.7**

As can be seen, modifying the separation between the windows appears to have a much more ambiguous effect. On the DEVSET in the EOV condition, for which the end-to-end prediction system was originally sanitized [3] (prior to the modification of the filterbank structure in [4]), the assymmetrical windows with $t_{sep} = 0.008$ s, $t_{int} = 0.008$ s, and $t_{exp} = 0.012$ appear to yield the best performance on the majority of measures. However, for the EOT condition, which is the primary condition of interest, providing h_L and h_R with symmertrical support of $t_{int} = t_{ext} = 0.010$ s, and separating them by $t_{sep} = 0.012$ s, appears to lead to better performance with our product-of-likelihoods classifier ("prod"). It is also the case that at these parameter values, individual hyperplanes induced by the 20 HMMs perform better on average, as evidenced by higher "mean" (and "min") discrimination.

In the EOV condition, these observations do not appear to generalize to the EvalSet, where the best "prod" performance lies at the intermediate parameter 3-tuple of $t_{sep} = 0.010$ s, $t_{int} = 0.009$ s, and $t_{ext} = 0.011$ s. We note that although the "min" discrimination consistently drops as t_{int} approaches and then surpasses t_{ext}, "mean" discrimination generally rises. We also observe a general increase in "mean" discrimination in the EOT condition for the EvalSet. However, most surprising is the very steep rise in "min" discrimination; when $t_{sep} = 0.014$ s, $t_{int} = 0.011$ s, and $t_{exp} = 0.009$ s, "min" and "max" discrimination across all 100 hyperplanes are very similar, yielding the lowest variance multiple HMMs trained on the same data of any of our experiments.

7   Conclusion & Future Directions

We have provided a detailed account of the computation of the fundamental frequency variation (FFV) spectrum, within the speaker-change prediction component of a spontaneous spoken dialogue system. Our description has focused on an aspect which has received little attention in our previous work, namely the design decisions surrounding spectral estimation for the left and right halves of each analysis window.

Our experiments have shown that asymmetrical h_L and h_R windows, skewed towards each other to minimize overlap in temporal support, do not lead to inferior performance relative to symmetric windows, in spite of their well-known poorer frequency resolution. Trends observed on the development data are also broadly present in evaluation data, and abandoning the symmetry criterion leads to relative improvements in ROC discrimination of 35–49% for the DevSet and 10-22% for the EvalSet.

However, our attempts to increase the temporal separation between the h_L and h_R windows have revealed different trends, and different optimal parameter values, for the two data sets. A notable exception is that under the target EOT condition, moving the windows further apart, at the expense of skewing them away from each other, appears to increase the robustness of our classifier. This is due to lower dependence on HMM initialization.

The experimental results presented suggest that further exploration of optimal window separation should proceed only in the context of a larger data set. This approach is also expected to level the absolute performance differences between the current DevSet and EvalSet.

8   Acknowledgments

References

[1] P. Heeman, "Modeling speech repairs and intonational phrasing to improve speech recognition", *Proceedings of IEEE/ACL Workshop on Automatic Speech Recognition and Understanding*, Keystone CO, USA (1999).

[2] P. Helgason, "SMTC — A Swedish Map Task Corpus", *Working Paper 52: Proceedings of Fonetik*, Lund, Sweden (2006).

[3] K. Laskowski, J. Edlund, M. Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems", *Proceedings of ICASSP 2008*, Las Vegas NV, USA (2008).

[4] K. Laskowski, J. Edlund, M. Heldner, "Learning prosodic sequences using the fundamental frequency variation spectrum", *Proceedings of Speech Prosody*, Campinas, Brazil (2008).

[5] D. Litman, "Classifying cue phrases in text and speech using machine learning", *Proceedings of 12th National Conference on Artificial Intelligence*, Seattle WA, USA, pp.806-813 (1994).

[6] D. Litman, M. Swerts, J. Hirschberg, "Characterizing and predicting corrections in spoken dialogue systems", *Computational Linguistics 32(3)*:417-438 (2006).

[7] G. Murray, S. Renals, M. Taboada, "Prosodic correlates of rhetorical relations", *Proceedings of the HLT/NAACL Workshop "Analyzing Conversations in Text and Speech"*, pp.1-7 (2006).

[8] A. Norwine, O. Murphy, "Characteristic time intervals in telephonic conversation", *Bell System Technical Journal 17*:281-291 (1938).

[9] V. Rangarajan, S. Bangalore, S. Narayanan, "Exploiting prosodic features for dialog act tagging in a discriminative modeling framework", *Proceedings of Interspeech*, Antwerpen, Belgium (2007).

[10] N. Ward, W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese", *Journal of Pragmatics 32*:1177-1207 (2000).