INTERSPEECH 2009 BRIGHTON

A General-Purpose 32 ms Prosodic Vector for Hidden Markov Modeling

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

Prosody plays a central role in conversation, making it important for speech technologies to model. Unfortunately, the application of standard modeling techniques to the acoustics of prosody has been hindered by difficulties in modeling intonation. In this work, we explore the suitability of the recently introduced fundamental frequency variation (FFV) spectrum as a candidate general representation of tone. Experiments on 4 tasks demonstrate that FFV features are complimentary to other acoustic measures of prosody and that hidden Markov models offer a suitable modeling paradigm. Proposed improvements yield a 35% relative decrease in error on unseen data and simultaneously reduce time complexity by a factor of five. The resulting representation is sufficiently mature for general deployment in a broad range of automatic speech processing applications.

1. Introduction

Prosody plays a crucial role in conversation, as it is associated with the structuring of speech as well as with speaker attitude and intention. Computational approaches to measuring prosodic phenomena are therefore important in automatic speech processing systems [1], if such systems are to behave in a manner consistent with human expectations of their interlocutors [2]. Although many frame-level features have been proposed which correlate with loudness, speaking rate, voice quality, and rhythm, those which correlate with pitch typically require long observation times for robustness. Pitch is also strongly speaker-specific and, for many tasks, requires additional estimation of normalization parameters [3, 4]. These aspects, and the discontinuity of pitch at the edges of voicing, make intonation difficult to model for individual frames.

To address this problem, an instantaneous and continuous representation of fundamental frequency variation (FFV) has been proposed [5, 6]. To date, it has been shown to be useful for speaker-change prediction [5] and floor mechanism detection [7], in anechoic-chamber and non-anechoic-chamber recordings, respectively. In addition, FFV bias has been shown to be speaker-discriminative in the same ways that Mel-cepstral features are, given single-state Gaussian mixture models [8]. In spite of advances in these application areas, FFV computation is approximately realtime, making it relatively expensive.

In the current work, we explore FFV features to detect dialog acts (DAs) known as floor holders and holds [7], and ask the following four questions (in Sections 4, 6, 7, and 8, respectively), ultimately answering them in the affirmative:

1. *Is feature space rotation necessary for tasks related to inference of dialog structure, in non-anechoic recordings?*
2. *Do FFV features yield improved performance when combined in feature space with correlates of loudness, voicing, and speaking rate?*
3. *Can the computation time be reduced at minimal cost to task accuracy?*
4. *Do standard higher-complexity acoustic modeling techniques apply to FFV features?*

Improvements resulting from our investigations yield consistent and significant gains, cumulatively reducing an agglomerated development set error measure on four related tasks by 11.1% absolute, or 39.0% relative. The corresponding reduction of the same error measure on an equally large unseen data set is 10.3% absolute or 35.4% relative.

2. Data

The data used in this work is drawn from the ICSI Meeting Corpus [9] and its associated MRDA dialog act annotations [10][1]. To our knowledge, it is the largest publicly available corpus of naturally-occurring multiparty conversation, consisting of longitudinal collections of meetings by several groups, and amounting to over 66 hours of meeting time. As defined in its 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, development exemplars from DEVSET, and unseen testing exemplars from EVALSET.

We consider 4 different but related binary classification tasks for which data is drawn separately:

Task 1A classification of the first 500 ms of each talkspurt as implementing a floor holder (or hold) vs. another DA type;

Task 2A classification of the last 500 ms of each talkspurt as implementing a floor holder (or hold) vs. another DA type;

Task 1B classification of the first 500 ms of each DA beginning in mid-talkspurt as implementing a floor holder vs. another DA type; and

Task 2B classification of the last 500 ms of each DA ending in mid-talkspurt as implementing a floor holder vs. another DA type.

Talkspurts, as used here, are contiguous intervals of speech and are obtained using forced alignment of human-transcribed words. Tasks 1A and 2A assume that only this segmentation is available. Tasks 1B and 2B, in contrast, assume that the speech

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

	TRAINSET	DEVSET	EVALSET
Task 1A	5000	1000	1000
Task 2A	5000	1000	1000
Task 1B	1200	240	240
Task 2B	750	180	180

Table 1: Number of instances of both classes in each binary classification task, per data set.

stream has been DA-segmented (but not DA-classified), which may or may not be the case in a fully automatic setting.

We note that by frequency of occurrence, floor holders and holds account for only a small proportion of DAs produced during a conversation. In this work, we consider each classification task on a *balanced prior* set, meaning that the number of exemplars for both classes is bound by the number of corresponding DAs in the minority class. These numbers are given in Table 1; exemplars were drawn randomly from each data set. The resulting set of exemplars is the same as used in [7].

3. Baseline

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 [5, 7]; 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. We note that frame energy is normalized out of this representation.

We oversample $g(\rho)$ at discrete equi-spaced intervals of ρ, and then pass the resulting vector through a filterbank whose design was motivated by psychoacoustic studies [11] (and whose central 5 filters are shown in Figure 1(a)). This leads to the 7-element representation of [5, 7].

For each binary classification task, we estimate 10 hidden Markov models (HMMs) $\mathcal{M}_c$ per class c over sequences of feature vectors using maximum likelihood expectation-maximization[2] (EM), from training material belonging to that class. As in [5], we use models of 4 fully-connected states and a single 7-dimensional Gaussian for the emission probability of each state. Automatic classification is performed by using the ratio of the average log likelihood (LL), over 10 HMM models, of a candidate sequence. Varying the ratio threshold allows for easy construction of receiver operating curves (ROC).

In scoring our systems, we use two agglomerate error measures over the four proposed tasks. First, we report the classification error when the LL ratio threshold is zero; the agglomerate measure is weighted by the number of training exemplars in each task. Because in natural settings priors over the phenomena of interest are quite skewed, we also report the ROC discrimination (the area below the ROC curve). Agglomerate ROC discrimination is also weighted by the number of exemplars in TRAINSET. Both measures for the baseline system are shown in line 1 of Table 2, and are identical to those in [7].

[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).

4. Feature Space Rotation

Our experience from [5] has been that, for speaker-change prediction, scaling and/or rotation of the feature space significantly improves classification using the baseline model. Trends are similar for the current task, as shown in Table 2; (2a) represents mean subtraction and variance scaling (Z-transform), with parameters inferred from TRAINSET, while (2b) shows results following a global PCA transform (also inferred from TRAINSET). For Z-transformed data, we observe a reduction of classification error of 3.0% absolute and of ROC discrimination error of 3.3% absolute, relative to the raw feature baseline.

System		Acc	ROC
1	Baseline	64.8	71.5
2a	Z-Transform	65.7	73.8
2b	PCA Rotation	67.8	74.8
3	Quadratic Fit	68.9	76.7
4a	Auxiliary Features	64.8	71.3
4b	Combination	69.6	77.6
5a	Exclusion of Extremity Filters	69.1	77.0
5b	Improvement of Extremity Filters	70.5	78.9
6	4 states, 2 Gaussians	71.7	80.3
7a	8 states, 1 Gaussian	71.3	79.4
7b	8 states, 2 Gaussians	73.0	81.5
7c	8 states, 3 Gaussians	**73.3**	**82.4**

Table 2: Accuracy in % for a log-average-likelihood-ratio threshold of zero and ROC discrimination in % on DEVSET.

5. A Modified Filterbank

Our attempts to render the FFV features more robust have led to a modified filterbank structure, and were motivated by the parabolic shape of the locus of responses in the 5 central filters (cf. [7]), which we refer to as $\mathbf{G}_5$. We have replaced $\mathbf{G}_5$ with its least-mean-squares (LMS) quadratic approximation

$$\tilde{\mathbf{G}}_5 \;=\; \begin{bmatrix} (-2)^2 & -2 & 1 \\ (-1)^2 & -1 & 1 \\ (0)^2 & 0 & 1 \\ (+1)^2 & +1 & 1 \\ (+2)^2 & +2 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \qquad (1)$$

for the LMS parabola described by $y = ax^2 + bx + c$; we refer to the matrix in Equation 1 as $\mathbf{X}$. This defines a rotation of $\mathbf{G}_5$,

$$\tilde{\mathbf{G}}_5 \;=\; \left(\mathbf{X} \cdot \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \cdot \mathbf{X}^T\right) \mathbf{G}_5 \;, \qquad (2)$$

which, when composed with the baseline filterbank, effectively induces a new filterbank structure; both are shown in Figure 1. As shown in Table 2 (line 3), the filterbank modification yields a 1.1% absolute and a 1.9% absolute reduction of the classification error and the ROC discrimination error, respectively. Although the new filterbank exhibits unexpected response characteristics (e.g. broader support of filters corresponding to slow pitch change), these consistent reductions of an agglomerate error (and almost all individual task errors in Table 3) suggest that it nevertheless better approximates human perception.

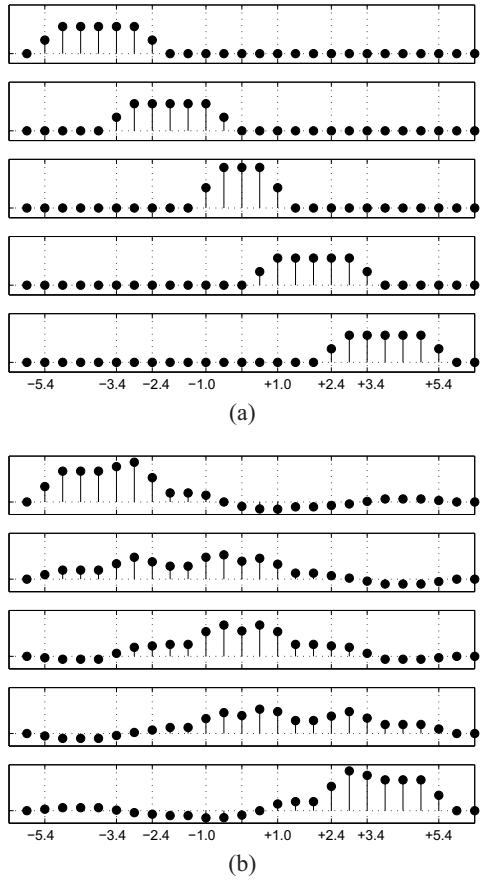

Figure 1: The 5 central filters of the original (a) and of the modified (b) filterbank. The x-axis is in semitones per 8 ms.

6. Combination with Other Features

As mentioned in the introduction, an important question is whether the FFV spectrum combines with other well-understood prosodic features to yield further improvements. As suggested in [7], it appears to also capture differences in speaking rate; in [8], when applied to the task of speaker identification, it was shown to degrade performance when combined with MFCC features at the feature level (but led to improvements when combined in model-space).

We propose an auxiliary vector of 5 relatively standard features, namely: overall log-energy above 300 Hz as a correlate of loudness, a log-energy difference between the two windows used to compute the FFV spectrum (as a correlate of change in loudness), the normalized height of the first autocorrelation peak (which is used as an indicator of the probability of voicing), and the cosine similarity measure between the Mel-spectra for the two FFV windows, in both the raw and log domain. The latter are energy-normalized measures of spectral flux, and correlate with speaking rate. Table 2 shows that the performance of these features is significantly above random guessing on this task but lower than that of the FFV spectrum (at line 3) by 4.1% absolute and 5.4% absolute for classification and ROC discrimination errors, respectively; feature-space combination with FFV spectrum features leads to improvements of 0.7% and 0.9%, respectively, relative to FFV features alone. We applied a global PCA transform to all 12 features in the combined feature space, and have not attempted other combinations.

7. Runtime Improvements

Next, we address the problem entailed by the two extremity filters in the baseline filterbank, whose combined support is 256 points, requiring us to scale $\mathbf{F}_L$ and $\mathbf{F}_R$ 128 times each (followed by 256 256-point dot products). This is significantly in excess of the 23 points comprising the support of the 5 central filters, $\mathbf{G}_5$, as shown in Figure 2. These two filters appear important; during voiced speech, they capture local minima corresponding to ± 1-octave errors and thereby play an implicit normalization role for $\mathbf{G}_5$ (via PCA). Eliminating them leads to degraded performance, shown in line 5a of Table 2 relative to line 4b.

Instead, we propose an improved design by fixing the support of the extremity filters to 23 points each, centered on the locations where we expect the largest magnitude ± 1-octave error. Surprisingly, this modification not only reduces time complexity by a factor of more than five (from 256 to 2×23 dot dilations and products), it also leads to a reduction of both classification error and ROC discrimination error, of 0.9% and 1.3% respectively.

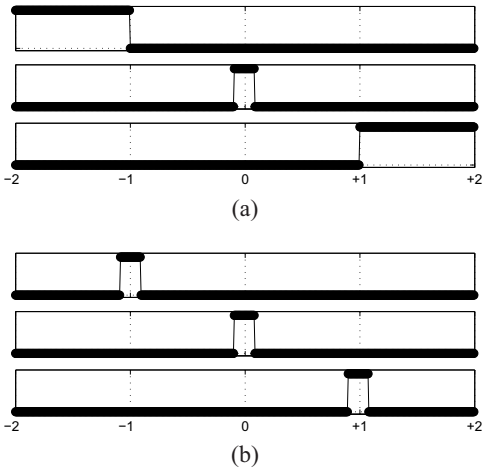

Figure 2: Left extremity filter, temporal support of the 5 central filters, and right extremity filter of the original (a) and of the modified (b) filterbank. The x-axis is in octaves per 8 ms.

8. Standard GMM Modeling

Finally, we sample several points in model complexity to verify the applicability of techniques which have long become standard for acoustic models in automatic speech processing applications. These are shown in lines 6 and 7a-c of Table 2. We stress that our goal is demonstrative and not intended to find a global optimum; these parameters should be optimized on a corpus of continuous speech rather than on isolated DA start and end snippets. However, the results clearly show improvement as both the number of states per HMM and components per GMM are increased. The configuration with highest complexity achieves reductions of 2.8% and 3.5% for the classification and ROC discrimination errors, respectively.

System	Task 1A		Task 2A		Task 1B		Task 2B	
	DEVSET	EVALSET	DEVSET	EVALSET	DEVSET	EVALSET	DEVSET	EVALSET
1	74.9	74.1	70.1	67.0	67.3	64.7	76.0	77.7
2b	75.8	76.0	74.3	70.8	68.4	66.4	79.3	82.6
3	78.3	77.9	73.4	71.3	70.8	68.2	81.0	83.3
4b	79.0	79.6	75.7	75.0	76.2	72.4	82.8	83.6
5b	79.2	80.2	75.8	75.5	78.5	74.5	81.2	82.7
7a	80.3	82.0	77.4	75.8	78.9	75.5	82.7	81.8
7b	81.4	83.6	78.6	**78.7**	**83.2**	**77.3**	81.9	83.1
7c	**82.2**	**83.9**	**79.8**	78.3	82.8	75.3	**85.1**	**84.7**

Table 3: ROC discrimination in % on DEVSET and EVALSET, using the single best pair of density models for classification as selected using DEVSET, for the four individual tasks in our study. Best performance per column shown in bold.

9. Generalization to Unseen EVALSET Data

In this final experimental section, we investigate how the improvements observed on DEVSET generalize to data not seen during development. To facilitate experiments, we select that pair of HMMs, out of the 10 trained for each binary class in each task, which yields the highest ROC discrimination on DEVSET. We then apply that pair to classify instances in EVALSET. The results are shown individually for all 4 tasks for both DEVSET and EVALSET, in Table 3.

As can be seen, with very few exceptions, the improvements shown in Table 3 yield consistent improvements on each task. Where exceptions do arise, they are limited to only one of DEVSET or EVALSET (such as line 3 for task 2A), or they occur for tasks 1B and/or 2B for which the amount of training material was relatively small (cf. Table 1).

Overall, our improvements appear to reduce baseline performance differences across the four tasks. Task 2A, corresponding to end-of-talkspurt detection of floor holders and holds, continues to prove more difficult than its beginning-of-talkspurt counterpart. The most dramatic improvement can be seen for task 1B (mid-talkspurt DA-terminal detection of floor holders) on DEVSET, but the proposed modifications have a much smaller impact on EVALSET data for this task. The relatively large increases in ROC discrimination observed in line 4b for tasks 2A and 1B suggest that our auxilary features play a large role for these two tasks. This may corroborate our findings in [7], and indicates that although the FFV spectrum seems to also capture some speaking rate effects, features specific for that purpose may exhibit significant complimentarity.

10. Conclusions and Future Directions

We have performed several important modifications to the FFV spectrum and benchmarked their performance, both individually and collectively, on 4 tasks related to floor control in naturally-occurring, multi-party speech. Of the proposed modifications, the most important include: a demonstration of complimentarity with features correlated with loudness, voicing, and speaking rate; a demonstration of applicability of a standard acoustic model paradigm; and a reduction of processing time by a factor of five. We have also proposed improvements of a structural nature, involving improved filterbank design and feature rotation. All modifications show consistent improvements when averaged across the 4 tasks explored, on both development and unseen test data. Absolute reductions of ROC discrimination error for the two data sets were shown to be 11.1% and 10.3% respectively, comprising relative reductions of 39.0% and 35.4%.

Our results suggest that, in its current state, the proposed FFV representation of instantaneous intonation is directly deployable in the acoustic space of a variety of speech processing applications, in conjunction with standard feature extraction such as that of MFCCs. Potential applications include additional discrimination for automatic speech recognition of tonal languages, as well as alternate phone and/or word models implementing competing dialog act productions.

11. Acknowledgments

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

12. References

[1] Sagisaka, Y., Campbell, N. and Higuchi, N. (eds), *Computing Prosody*, Springer, 1997.

[2] Edlund, J., Gustafson, J., Heldner, M. and Hjalmarsson, A., "Towards human-like spoken dialogue systems", in *Speech Communication*, **50**(8-9):630–645, 2008.

[3] Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E. and Niemann, H., "Boiling down prosody for the classification of boundaries and accents in German and English", in *Proc. EUROSPEECH*, 2001.

[4] Ferrer, L., Shriberg, E. and Stolcke, A., "Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody", in *Proc. ICSLP*, 2061–2064, 2002.

[5] Laskowski, K., Edlund, J. and Heldner, M., "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems", in *Proc. ICASSP*, pp. 5041–5044, 2008.

[6] Martin, P., "A fundamental frequency estimator by crosscorrelation of adjacent spectra", in *Proc. SPEECH PROSODY*, 2008.

[7] Laskowski, K., Heldner, M. and Edlund, J., "Exploring the prosody of floor mechanisms in English using the fundamental frequency variation spectrum", in *Proc. EUSIPCO*, 2009.

[8] Laskowski, K. and Jin, Q., "Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum", in *Proc. ICASSP*, 4541-4544, 2009.

[9] Janin, A. et al, "The ICSI Meeting Corpus", in *Proc. ICASSP*, 364-367, 2003.

[10] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J. and Carvey, H., "The ICSI MRDA Corpus", in *Proc. SIGdial*, 97-100, 2004.

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