MODELING INSTANTANEOUS INTONATION FOR SPEAKER IDENTIFICATION USING THE FUNDAMENTAL FREQUENCY VARIATION SPECTRUM

Kornel Laskowski [1,2] *and Qin Jin* [1]

[1] interACT, Carnegie Mellon University, Pittsburgh PA, USA
[2] Cognitive Systems Lab, Universität Karlsruhe, Karlsruhe, Germany

ABSTRACT

In recent years, the field of automatic speaker identification has begun to exploit high-level sources of speaker-discriminative information, in addition to traditional models of spectral shape. These sources include pronunciation models, prosodic dynamics, pitch, pause, and duration features, phone streams, and conversational interaction. As part of this broader thrust, we explore a new frame-level vector representation of the instantaneous change in fundamental frequency, known as fundamental frequency variation (FFV). The FFV spectrum consists of 7 continuous coefficients, and can be directly modeled in a standard Gaussian mixture model (GMM) framework. Our experiments indicate that FFV features contain useful information for discriminating among speakers, and that model-space combination of FFV and cepstral features outperforms cepstral features alone. In particular, our results on 16kHz Wall Street Journal data show relative reductions in error rate of 54% and 40% for female and male speakers, respectively.

Index Terms— Speaker identification, Intonation, Fundamental frequency.

1. INTRODUCTION

Automatic speaker identification (SID) is the inference by machine of speaker identity, given a recorded utterance. The task has traditionally relied on the extraction of low-level spectral features, such as those in automatic speech recognition, and maximum likelihood classification using continuous density models. More recently, SID systems have been augmented with high-level features in an effort to exploit what is known about speaker identification by humans [1]. This includes prosodic features and, in particular, features descriptive of fundamental frequency (F0) and its variation.

F0 features have been applied to the SID task in one of three main ways. Most commonly, global utterance-level statistics, such as mean and standard deviation, are estimated and compared between two utterances [2]. However, such statistics do not capture the shape of the F0 trajectory in time, a limitation which has been addressed in part through the inclusion of dynamic features in the feature vector [3]. A second approach to modeling F0 for SID aims to explicitly represent the F0 trajectory in time. Pitch contours between any two renderings of the same lexical content can be compared using dynamic time warping [4]. This approach is limited to text-dependent speaker recognition applications. Its extension to text-independent applications [5], comprising the third approach we mention, relies on the availability of a speech recognition system and requires a considerable amount of training data. In theory, when these requirements are met, the approach allows for the inference of conditional F0 feature densities, given other features such as energy trajectories or specific lexical contexts.

In this paper, we apply to the problem of speaker identification a frame-level vector representation of the *instantaneous change* in F0, known as the fundamental frequency variation (FFV) spectrum [6] introduced in [7]. Unlike F0, the FFV spectrum remains well-defined in the absence of voicing, and eliminates the need to localize a unique peak corresponding to the fundamental frequency, a process which is prone to error. The experiments we present suggest that these properties of the FFV spectrum recommend it for standard Gaussian mixture modeling (GMM). We successfully show that, for speaker identification, FFV information is complementary to that in standard frame-level Mel-frequency cepstral coefficients (MFCCs). Model-based combination with a GMM-FFV system reduces the classification error rate of our baseline GMM-MFCC system by 40-54%. We note also that the current work is the first attempt to compute the FFV spectrum for a discriminative task using recordings made outside of the anechoic chamber.

The paper is organized as follows. Section 2 introduces the FFV representation, provides details necessary for its computation, and outlines how that computation differs from what has been described in the literature for a different discriminative task [8]. In Section 3, we present the baseline GMM-MFCC SID system and in Section 4 describe the datasets used in this work. Experiments quantifying the performance of the baseline and that of the incremental gain from FFV features are presented in Sections 5 and 6, respectively. We conclude in Section 7 with a summary of our findings and suggest several avenues for future work.

2. FUNDAMENTAL FREQUENCY VARIATION

The FFV coefficients are computed from a single frame of audio, much like the Mel-frequency cepstrum. The computation is detailed in [7, 6, 8]; here, we provide only a brief account:

1. Compute a 512-point FFT over the left half of the analysis frame, and similarly for the right half of the analysis frame (henceforth $\mathbf{F}_L$ and $\mathbf{F}_R$, respectively); FFTs are computed using two 20ms-wide Hann windows whose peaks are $t_{sep} = 0.012$ seconds apart, as shown in Figure 1.
2. Compute the *vanishing point product* of $|\mathbf{F}_L|$ and $|\mathbf{F}_R|$, by dilating one or the other spectrum by a factor 2^ρ, for 512 discrete values of ρ, to yield the *FFV spectrum*

$$g_N^{\rho}\left(\rho\right) \;=\; \left\{ \begin{array}{ll} \dfrac{\sum|F_L[k]|\left|\tilde{F}_R^*\left(2^{+\rho}k\right)\right|}{\sqrt{\sum|F_L[k]|^2\sum\left|\tilde{F}_R^*\left(2^{+\rho}k\right)\right|^2}}\,, & \rho < 0 \\ \dfrac{\sum\left|\tilde{F}_L^*\left(2^{-\rho}k\right)\right|\left|F_R^*[k]\right|}{\sqrt{\sum\left|\tilde{F}_L^*\left(2^{-\rho}k\right)\right|^2\sum\left|F_R^*[k]\right|^2}}\,, & \rho \geq 0 \end{array} \right. . \qquad (1)$$

The dilated magnitude spectrum (either $|\tilde{\mathbf{F}}_L|$ or $|\tilde{\mathbf{F}}_R|$ in Equa-

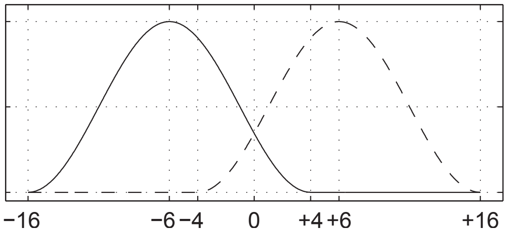

Fig. 1. Window functions used to produce estimates of $\mathbf{F}_L$ and $\mathbf{F}_R$, the frequency spectra of the left and right halves of each 32ms analysis frame (512 points at a sampling frequency of 16 kHz), respectively. x-axis coordinates are in ms.

tion 1) is linearly interpolated using

$$\beta_{\pm}^{\rho,k} = \left| \lceil 2^{\pm\rho} k \rceil - 2^{\pm\rho} k \right| \ , \tag{2}$$

$$\left| \tilde{F}\left(2^{\pm\rho} k\right) \right| = \beta_{\pm}^{\rho,k} \left| F\left[\lfloor 2^{\pm\rho} k \rfloor\right] \right| + \left(1 - \beta_{\pm}^{\rho,k}\right) \left| F\left[\lceil 2^{\pm\rho} k \rceil\right] \right| \quad . \tag{3}$$

The values ρ governing the dilation factor are given by their discrete counterparts $r \in [-256, 255]$, via

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

where $N{\equiv}512$. $t_{sep}^{ref}{\equiv}0.008$ s is the original value for the inter-peak spacing in Figure 1, taken from [7].

3. Compress the FFV spectrum using a 7-filter filterbank, as described in [6]; the design of the filters is perceptually motivated and a function of t_{sep}^{ref}, with one filter for unchanging pitch (henceforth FFV_0), one for each of slowly rising and slowly falling pitch (FFV_{+1} and FFV_{-1}, respectively), one for each of quickly rising and quickly falling pitch (FFV_{+2} and FFV_{-2}, respectively), and two filters for indeterminate pitch ($FFV_{\pm 3}$).
4. Decorrelate the compressed FFV spectrum using a normalizing and/or decorrelating transform, such as the Z-transform, the Karhunen-Loéwe transform, the sine transform, or the cosine transform. In the current work, we use a single, global Karhunen-Loéwe transform, inferred from the training data.

We note that the above-described procedure deviates slightly from that used for predicting speaker change in dialog systems [7, 6, 8]. In particular, for Step 1, our earlier work relied on windows which were placed closer together but skewed away from each other, and which did not attain zero at ±16 ms. Experiments presented in [8] showed that on held-out data, the two windows shown in Figure 1 are better for speaker change prediction than those originally proposed [7]; we have performed only a limited subset of sanity experiments on the current task to fully tune the window shapes to the current task. Steps 2, 3, and 4 are identical to those described in [6], where the current settings were shown to be optimal for speaker change prediction. Step 3, whose filterbank calls for the evaluation of Equation 1 over only a subdomain of ρ, offers significant scope for processing-time reduction.

We note that an alternative account of computing the FFV spectrum, not in the context of a discriminative task but with the aim of improving F0 estimation, can be found in [9].

3. A BASELINE GMM-MFCC SID SYSTEM

The baseline SID system used in this work is shown in Figure 2. It operates in one of three modes: (1) universal background model (UBM) training; (2) target speaker enrollment; and (3) testing.

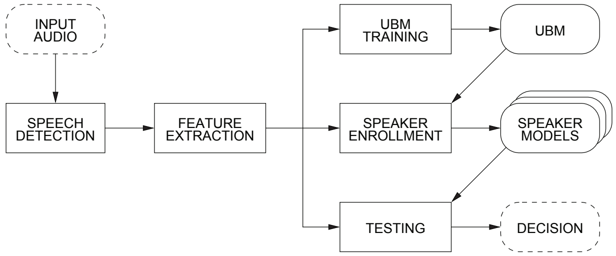

Fig. 2. GMM-based Speaker Identification System

All three modes of operation rely on identical feature extraction, which frames the signal into 32 ms windows every 8 ms. We have chosen this framing policy, developed for speaker change prediction using the FFV spectrum [7], in order to facilitate the feature-space combination of MFCC and FFV features; previous experiments have shown that MFCC baseline system performance is not very sensitive to changes in frame size (in the range 20-50 ms) and frame step (in the range $1/6$ to $1/3$ of the frame size). Frames whose energy is too low to be considered speaker-discriminative are excluded from subsequent processing. From each surviving frame, the first 13 Mel-frequency cepstral coefficients are computed and normalized using cepstral mean substraction (CMS), yielding the feature vector

$$\text{MFCC13} \equiv \{MFCC_0, MFCC_1, \ldots, MFCC_{12}\} \ . \tag{5}$$

First- and second-order differences are not used, as they have not been shown to help on this data.

We model sequences of feature vectors, assumed independent and identically distributed, with Gaussian mixture models, the most frequently and successfully employed density estimators in speaker identification [10, 11]. Every target speaker's model consists of 4096 components, each parameterized by its a priori probability, mean vector, and diagonal covariance matrix. We use 4096 Gaussians here because we anticipate 5 minutes of training data, and have successfully used 1024 Gaussians in cases where the training data was only 1-2 minutes in duration. Target speaker enrollment is preceded by UBM parameter inference [11], which is accomplished via the expectation-maximization (EM) algorithm using a large corpus of speech from non-target speakers. Target speaker model means are then adapted away from the UBM via maximum a posteriori (MAP) estimation, using only the target speaker's speech.

Testing proceeds by applying the same feature processing as for model training. The observed sequence of feature vectors is then scored by each speaker's model. The system hypothesizes that speaker whose model best accounts for the observed sequence. Performance is assessed using identification accuracy.

4. DATA

Experiments described in the current work use data drawn from the LDC CSR-I (WSJ0) [12] and LDC CSR-II (WSJ1) [13] corpora. Speech snippets consist primarily of read sentences from the Wall

Street Journal, but also include some spontaneously produced utterances. They were selected from files in the published corpora which had a `.wv1` extension (indicating a Sennheiser HMD414 close-talk head-mounted microphone).

For each target speaker, TRAINSET and TESTSET contributions were constructed by including utterances until there were at least 5 minutes of speech data per speaker for training, and at least 3 minutes, in 3 trials of 1 minute each, for testing. In total, we identified enough speech data for 95 male speakers and 102 female speakers. The total number of test trials was 285 and 306, for our male and female speakers, respectively. Speech from the remaining speakers (approximately 70 hours) was placed in UBMSET for UBM training.

Our experiments are intentionally limited to same-gender speaker identification, as our gender classifier can detect gender in this data with 100% accuracy. The database as described was designed for several ongoing experiments in our lab, and was motivated by a need for wideband 16kHz audio, a large number of potential speakers, a large amount of speech per speaker, and the availability of relatively clean transcripts. The corpus represents an intermediate step towards our ultimate goal of evaluating the new FFV features for far-field speaker identification.

5. BASELINE SYSTEM PERFORMANCE

In Table 1, we show the performance of the baseline SID system (MFCC13), together with that of several contrastive systems. In particular, we show the classification accuracy achieved by a system relying only on higher-order MFCC coefficients,

$$\text{MFCC7} \quad \equiv \quad \{MFCC_{13}, \ldots, MFCC_{19}\} \ , \tag{6}$$

as well as that of various types of system combinations with the MFCC13 baseline. This includes feature-level combination (⊕) and model-level or score-level combination (⊗). We have chosen the 7 subsequent MFCC coefficients to make system combinations comparable in complexity to those involving the FFV7 features.

Combination at the feature level consists of concatenating the MFCC13 feature vector with the additional feature vector, e.g.,

$$\text{MFCC13} \oplus \text{MFCC7} \quad = \quad \{MFCC_0, \ldots, MFCC_{19}\} \ , \tag{7}$$

i.e., a new 20-dimensional feature vector. Assessment of performance of this combination type requires the training of new models. For simplicity, in doing so, we do not modify the number of Gaussians from that employed for the 13-feature MFCC13 baseline.

In contrast, score-level combination does not require new model training. Identification is performed using a new score

$$\begin{aligned} &\log P(\,\mathbf{x}\,|\,\mathcal{M}_1^k \otimes \mathcal{M}_2^k\,) \\ &\qquad \doteq \ \lambda_1 \log P(\,\mathbf{x}\,|\,\mathcal{M}_1^k\,) + \lambda_2 \log P(\,\mathbf{x}\,|\,\mathcal{M}_2^k\,) \ , \end{aligned} \tag{8}$$

where $\mathbf{x}$ represents an utterance's sequence of feature vectors and $\mathcal{M}_i^k$ is the model for feature set i and speaker k. We split TESTSET into three subsets, consisting of all speakers' first trials, second trials, and third trials, and estimate the weights for each subset based on individual model $\{\mathcal{M}_i^k\}$ accuracy A_i on the other two subsets, in round-robin fashion, using $\lambda_i = A_i / \sum_j A_j$. Overall TESTSET accuracy is obtained by summing along the diagonal of the sum of the three subset confusion matrices.

As can be seen from Table 1, the MFCC7 system by itself is significantly weaker than the standard MFCC13 system, given identical model complexity. However, both feature-level combination (MFCC13 ⊕ MFCC7) and score-level combination (MFCC13 ⊗ MFCC7) result in accuracies which are at least as high as for MFCC13 alone, for both female and male speakers. Both of these combinations will be treated as augmented contrast conditions, when assessing the incremental performance of the 7-feature FFV representation in Section 6.

System	Female	Male
MFCC13	82.0	92.3
MFCC7	43.8	68.8
PCA(MFCC13)	**84.3**	91.2
PCA(MFCC7)	44.4	65.2
MFCC13 ⊕ MFCC7	**87.9**	**94.0**
PCA(MFCC13) ⊕ MFCC7	**87.5**	**94.0**
MFCC13 ⊕ PCA(MFCC7)	**88.9**	**94.0**
PCA(MFCC13) ⊕ PCA(MFCC7)	**90.2**	**94.0**
PCA(MFCC13 ⊕ MFCC7)	**87.5**	**94.0**
MFCC13 ⊗ MFCC7	**86.3**	**92.6**
PCA(MFCC133) ⊗ MFCC7	**87.3**	92.3
MFCC13 ⊗ PCA(MFCC7)	**89.2**	**93.7**
PCA(MFCC13) ⊗ PCA(MFCC7)	**86.9**	**93.0**

Table 1. Identification accuracies (in %) using several combination types of standard MFCC features. Numbers in bold represent combinations which outperform the MFCC13 baseline.

Table 1 also shows the effect of applying a global Karhunen-Loéwe transform, obtained via principal component analysis (PCA) of the entirety of the training data (TRAINSET and UBMSET). Interestingly, PCA leads to improved performance for female speakers but slightly degraded performance for male speakers, for both MFCC13 and MFCC7 features when these are used alone. When combining MFCC13 and MFCC7 features at the feature-level, the global PCA transform has no effect on male speaker classification accuracy. However, for female speakers, applying PCA to MFCC13 only, or to the feature-level combination MFCC13 ⊕ MFCC7, reduces accuracy to 87.5%. This suggests that PCA must be applied to the MFCC7 features independently of the MFCC13 features for female speakers. Female speaker classification accuracy is highest (90.2%) when feature-level combination is applied following separate PCA transforms for both MFCC13 and MFCC7.

Finally, score-level combination leads to accuracies which are at least as high as those for the baseline MFCC13 system, but never outperforms the best feature-level combination. For both genders, the best score-level combination accuracy is 0.3-1.0% absolute lower than the best feature-level combination accuracy.

6. INCREMENTAL PERFORMANCE OF FFV FEATURES

In this section, we compare the incremental performance of FFV features to that of higher-order cepstral features, given an existing MFCC13 baseline; both the FFV7 and the MFCC7 feature vectors consist of 7 features. The results, involving the same types of system combination as in Section 5, are shown in Table 2.

We make 3 specific observations. First, FFV7 features alone offer better than chance performance, but are inferior to MFCC7 features (themselves inferior to MFCC13 features). Sphering the MFCC7 features has a small and ambiguous effect across both genders, while sphering the FFV7 features leads to large improvement; for female speakers, PCA renders FFV7 features significantly more powerful than MFCC7 features.

System	Female	Male
MFCC13	82.0	92.3
MFCC7	43.8	68.8
FFV7	27.8	45.3
PCA(MFCC13)	84.3	91.2
PCA(MFCC7)	44.4	65.2
PCA(FFV7)	**62.7**	64.2
MFCC13 ⊗ MFCC7	86.3	92.6
MFCC13 ⊗ FFV7	80.7	92.3
PCA(MFCC13) ⊗ MFCC7	87.3	92.3
PCA(MFCC13) ⊗ FFV7	84.6	**92.6**
MFCC13 ⊗ PCA(MFCC7)	89.2	93.7
MFCC13 ⊗ PCA(FFV7)	**91.8**	**95.4**
PCA(MFCC13) ⊗ PCA(MFCC7)	86.9	93.0
PCA(MFCC13) ⊗ PCA(FFV7)	**91.5**	**95.1**

Table 2. Identification accuracy (in %) using the model-space combination as in Table 1; performance of combinations involving MFCC13 and FFV7 is contrasted with those involving MFCC13 and MFCC7. Numbers in bold represent FFV7 combinations which outperform the identical combination with MFCC7 features.

Second, feature-level combination of MFCC13 and FFV7 features, regardless of the specifics of PCA application, always leads to performance which is inferior to that obtained by combining MFCC13 and MFCC7 features. Table 2 does not show these numbers due to space constraints.

Third, score-level combination with FFV7 leads to significant improvement over the MFCC13 baseline for both genders, but requires sphering of the FFV7 features. The best-performing combination type per gender reduces classification error with respect to the MFCC13 baseline by 54% for female speakers and by 40% for male speakers. In addition, when the incremental 7-feature system applies PCA, combination with FFV7 always outperforms that with MFCC7. Relative error rate reductions of 24-35% and 26-30% were observed for female and male speakers, respectively, by moving from combination with PCA(MFCC7) to that with PCA(FFV7).

We note in closing this section that score-level combination with either MFCC7 or FFV7 appears to also shrink the absolute gap between system performance on male and female data, suggesting that both MFCC7 and FFV7 offer improved modeling of the more widely separated harmonics characteristic of female speech. That score-level combination with FFV7 outperforms that with MFCC7 but feature-level combination exhibits the reverse behavior is somewhat unexpected, and is the focus of ongoing investigation.

7. CONCLUSIONS

We have explored the FFV spectrum, a continuous, vector-valued representation of variation in fundamental frequency, for speaker identification. Although initially proposed for predicting speaker changes in two-party dialogue, the features appear to discriminate among speakers, offering information which is complementary to MFCCs. In particular, on the data set studied, model-space combination with FFV features reduces classification errors by 40-54% relative to an MFCC baseline, for both female and male speakers. Our results suggest that modeling does not require large amounts of training data, as compared to other approaches intended to capture prosodic dynamics (e.g., [5]). Furthermore, because the FFV spectrum does not represent a speaker's absolute F0, standard modeling of absolute pitch is likely to lead to orthogonal improvement.

This work also represents the first attempt to infer FFV models via MAP adaptation, as well as the first successful attempt at applying FFV processing to audio recorded outside of an anechoic chamber. In future work, we aim to assess these features on farfield speech and to extend our experiments to larger datasets.

8. ACKNOWLEDGMENTS

We would like to thank Joe Campbell for his suggestion to apply the FFV spectrum to speaker identification, and Arthur Toth for preparing the database.

9. REFERENCES

[1] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID Project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP*, Hong Kong, China, 2003.

[2] M. Carey, E. Parris, H. Lloyd-Thomsa, and S. Bennett, "Robust prosodic features for speaker identification," in *Proc. ICSLP*, Philadelphia PA, USA, 1996, pp. 1800–1803.

[3] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 3189–3192.

[4] B. Atal, "Automatic speaker recognition based on pitch contours," in *J. ASA*, 1972, vol. 52, pp. 1687–1697.

[5] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. ICASSP*, Hong Kong, China, 2003, pp. 19–41.

[6] K. Laskowski, J. Edlund, and M. Heldner, "Learning prosodic sequences using the fundamental frequency variation spectrum," in *Proc. SPEECH PROSODY*, Campinas, Brazil, 2008.

[7] 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, 2008, pp. 5041–5044.

[8] 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, 2008, pp. 3305–3310.

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

[10] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," in *IEEE Trans. Speech and Audio Processing*, 1995, vol. 3, pp. 72–83.

[11] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, vol. 10, pp. 19–41.

[12] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," Linguistic Data Consortium, 2007, vol. LDC93S6A.

[13] "CSR-II (WSJ1) Complete," Linguistic Data Consortium, 1994, vol. LDC94S13A.