Modeling Prosody for Speaker Recognition: Why Estimating Pitch May Be a Red Herring

Kornel Laskowski and Qin Jin

Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
kornel@cs.cmu.edu qjin@cs.cmu.edu

Abstract

It has long been claimed that spectral envelope features outperform prosodic features on speaker recognition tasks. However, the reasons for such an arrangement are not entirely compelling. In the current work we present some evidence to challenge these claims. We propose that energy found at harmonically related frequencies encodes the acoustic correlates of variables which are typically referred to as prosodic, making harmonic energy summation highly relevant. Its frequent implementation for estimating pitch appears to have gone unnoticed by the speaker recognition community, because pitch estimators quite deliberately discard what they compute, retaining only the abscissa of a maximum. We argue that this latter step renders pitch estimation somewhat ill-suited to speaker recognition tasks. We present the detailed construction of a discrete transform, and a normalization, which are amenable to relatively laconic modeling. With this framework we achieve or exceed the performance of spectral envelope features in nearfield, matched-channel and matched-multisession conditions; performance improves following envelope destruction. We believe these results may have far-reaching consequences. For speech processing in a multitude of applications, they suggest that modeling the harmonic structure in the way we propose is at least as relevant as is modeling other aspects of the signal.

1. Introduction

Automatic speaker recognition is an important task in today's society, with diverse application in the management of access, security, and privacy. Modeling speaker characteristics can also benefit other speech technologies, whose direct aim is not the inference of speaker identity. It is therefore somewhat surprising, given this broad scope of impact, that the most stable and oft-cited baselines [1, 2] are those whose signal representation is both speaker-independent by design and optimized for a different task (namely speech recognition).

An analysis of acoustic features as used frequently in speech processing reveals that they fall into four categories, along two discrete dimensions. Along the first dimension, features may describe the instantaneous, within-frame characteristics of a signal; alternately, they may describe longer time-span variation across frames. Along the second dimension, features represent either the spectral envelope, broadly understood to encode the shape of the vocal tract, or aspects purported to characterize the glottal source or other excitation.

It is commonly understood that instantaneous features representing the spectral envelope are of immediate relevance to automatic speech recognition (ASR), and they are colloquially referred to as such. The term *prosodic features* has come to denote the other three categories, variation in which ASR systems attempt to normalize out. As might be expected, that variation has been shown to be useful in speaker recognition [3, 4, 5], but its utility by itself is relatively low and benefit stems from the fact that it is merely complementary to ASR features. Furthermore, prosodic features are known to be notoriously difficult to model [6] and to require large amounts of training material.

In this work, we focus on only a subset of these prosodic features, namely those that characterize within-frame aspects of the frequency magnitude spectrum. The most well-known among them is fundamental frequency (F_0), or pitch. We argue that the estimation of pitch, an $\arg\max$ operation in a transformed domain, does not serve the needs of speaker recognition. Those pitch errors which are considered most egregious may actually be just as speaker-discriminative as is pitch itself, if not more so. From a systems engineering perspective, the result is that the speaker discriminative information which pitch estimators may first compute, but then suppress or discard in the service of a better $\arg\max$ hypothesis, is never made available to downstream components. As one would expect, that information appears to be mostly unrecoverable even in the face of costly and arcane modeling efforts. Our goal is to reverse these seeming anomalies, at their source.

2. Pitch Estimation and Processing

2.1. Pitch Detection

A pitch detector is a within-frame 1-best $\arg\max$ locator, in some space which we refer to as the *transformed-domain*. Quite a few alternative transformed-domains have been proposed, and each is purported to have specific desirable properties making its $\arg\max$ a good estimate of the fundamental frequency. Transformed-domains which are popular include the comb filterbank energy spectrum [7, 8], the autocorrelation spectrum, the magnitude frequency domain, and the cepstral domain.

2.2. Error Types

Pitch detectors of course commit errors, and these fall into three categories [9]. First, pitch detectors may fail to identify whether a speech frame contains voicing. This error will then lead to an erroneous estimate of pitch (a type I error), or a failure to provide an estimate where one is sought (a type II error). Second, pitch detectors may commit what is known as an *octave* or a *suboctave* error, by misidentifying a rational multiple of the true F_0, namely $({}^{p}/_{q})\ F_0$ for $p \in \mathbb{N}, q \in \mathbb{N}, p \neq q$, as the F_0 estimate they seek. These errors are typically referred to as *gross errors*. Finally, pitch detectors may commit *fine errors*, of small magnitude relative to gross errors, for a variety of biomechanical, phonotactic, environmental, or measurement reasons.

Research in pitch detection has generally focused on eliminating the gross errors, since they are not unimodally distributed

and contribute disproportionally to cumulative error rates. Furthermore, they can frequently be identified visually, which is less true of both voicing errors and fine errors.

2.3. Pitch Tracking

Since pitch trajectories are physiologically constrained to be continuous between unvoiced occlusions [10], a useful approach to curtailing gross errors is to track the frame-level pitch estimate in time. This is frequently implemented via dynamic programming, and trackers are typically exposed to not only the 1-best but to N-best detector estimates per frame. The success of the continuity assumption has blurred the distinction between mere detectors and trackers, and currently no off-the-shelf detector exists which does not employ some form of tracking.

Although technological improvement to tracking in time is independent of improvement to frame-level pitch detection, the details of parameter tuning in dynamic programming implementations may be pitch-detector-specific. As such, there is a potential that, in lowering gross error rates, tuning may actually increase the occurrence and magnitude of fine errors (which may be speaker-specific).

2.4. Downstream Processing of Pitch Information

Despite the considerable investment in the complexity of pitch detection and tracking to reduce gross errors, downstream speech processing applications, such as speaker recognition systems, frequently find it important to further smooth the F_0 estimate sequence. Common approaches are median filtering and/or linearization; without a doubt these approaches eliminate potentially speaker-discriminative fine errors.

Following such gross pitch *correction* measures, systems designed to exploit pitch variation frequently compute a very large number of additional features, *derived* from the corrected trajectory estimate. These include derivatives, durations, separations; alignments with utterance, talkspurt, word or syllable boundaries; various normalizations of these features; and their statistics over some interval of fixed or variable duration. It is obvious that uncorrected underlying errors, distributed non-unimodally, would have serious deleterious impact on the utility of models of such features.

2.5. Desiderata

We now posit that improving pitch estimation may actually be somewhat orthogonal to speaker recognition and many other speech classification problems. This is because it seeks to eliminate the error type which contributes most to cumulative pitch detection error rates, namely octave errors. To see why this could be suboptimal for speaker recognition, consider the discrete-frequency domain in which the distance from $f = 0$ to each harmonic peak might be used as an *independent* feature characterizing each speaker. Assuming that measurement noise (and other sources of noise, such as frequency smearing or fine errors) does not scale linearly with frequency, and there is no reason to believe that it generally does, any two speakers with partially overlapping F_0 domains will have $(2 \cdot F_0)$ domains which are *less* overlapping, a property which is only the more relevant for the remaining $p > 2$ harmonics.

Even if this were not the case, and if estimation error did for some reason scale linearly with frequency, it is computationally unjustifiable to discard the entire transformed-domain signal except for its global $\arg\max$. It is conceptually tantamount to attempting to run a speech recognizer by discarding the entire frequency-domain signal except for the center frequency of e.g. the first formant.

It is important to note that prosodic speaker recognition systems frequently compute features other than instantaneous pitch, either from the pitch trajectory itself (as mentioned in Subsection 2.4) or via separate processing. Such other features may exhibit some correlation with the transformed-domain signal which has by then been discarded. The measurement of this correlation is beyond the scope of the present work.

3. Harmonic Structure Transform (HST)

The transform we present operates on the discrete-frequency energy domain, which we refer to as the FFT domain. This choice facilitates conceptualization of the phenomena most germane to pitch analysis from a perceptual standpoint, such as "fundamental frequency", "octave error", "harmonic", etc.

3.1. Preprocessing

The audio used in the current work was sampled at $f_s = 16$ kHz. We first apply a 32 ms Hann window every 8 ms; the Hann window appears to be much less sensitive to noise than the ubiquitous Hamming window [12]. Given these constraints, each frame of audio consists of 512 points. We transform this data into the frequency domain, yielding $N_f + 1 = 257$ distinct, equi-spaced, positive-half-frequency bins, spanning from 0 kHz to $f_s/2 = 8$ kHz. We assume these bins to be centered on the frequencies $f_c[j] = j \cdot \Delta f_c$, for $0 \leq j \leq N_f$, with $\Delta f_c \equiv f_s/N_f = 31.25$ Hz. We treat the bins as tesselating the frequency axis without overlap, such that the width of each bin is identically Δf_c.

From the complex spectra thus computed, we retain only the squared-magnitude response and discard the phase component, leading to input vectors $\mathbf{x} \in \mathbb{R}^{(N_f+1)}$.

3.2. Comb Filtering in the Frequency Domain

We now seek a linear filterbank transform, implemented as a matrix multiplication $\mathbf{H} \in \mathbb{R}^{(N_f+1)\times(N_h+1)}$, into a different domain which will become our feature space of vectors $\mathbf{y}$. We propose that this feature space be the *harmonic frequency* domain [11], representing energy distribution at *candidate fundamental frequencies and their harmonics*. This requires that the columns of $\mathbf{H}$ be *comb filters*.

A comb filter $h(f)$ is a digital FIR or IIR filter whose nulls in the frequency domain are located at integer multiples of some candidate fundamental frequency F_0 [8]. When F_0 is unknown, it may be estimated by minimizing the energy appearing at the filter's output, over a bank of such filters. The form of the filter, effectively a harmonic notch-stop filter, has an equivalent formulation in the continuous- time and frequency domains, as well as in the discrete- time and frequency domains; however, in the discrete frequency domain, it is more convenient to ignore spectral shaping considerations and to use a harmonic *notch-pass* counterpart [13, 9, 14].

We are at liberty to choose the candidate fundamental frequencies $\{f_h\}$, which we do in a most general manner by letting $f_h[i] = f_h^{min} + i \cdot \Delta f_h$, for $0 \leq i \leq N_h$. Setting $f_h^{min} \equiv 50$ Hz, $\Delta f_h \equiv 1$ Hz, and $N_h = 400$ produces a span from 50 Hz to 450 Hz, conservatively bracketing the overwhelming majority of observed fundamental frequency values in adult human speech.

For each i, $0 \leq i \leq N_h$, corresponding to a candidate fundamental frequency $f_h[i]$, we construct one comb filter

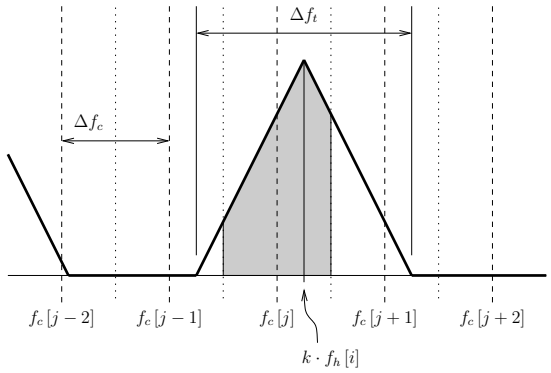


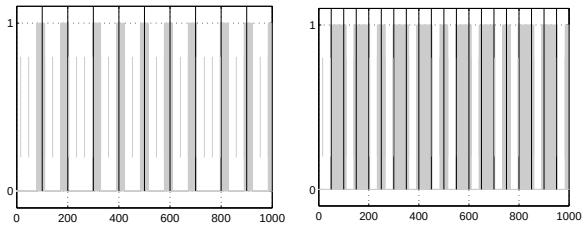
Figure 1: Riemann sampling of the kth idealized harmonic of the continous-frequency comb for fundamental frequency $f_h[i]$. Shown in gray is the area assigned, during filter construction, to the jth element of the corresponding discrete-frequency comb filter $\mathbf{h}_i$. Frequency shown along the x-axis; symbols as in the text.

$\mathbf{h}_i \in \mathbb{R}^{(N_f+1)}$. We first conceive of the comb filter in the continuous frequency domain, as the sum of $f_h[i]$-harmonically spaced Dirac delta functions convolved with a triangular peak shape $W(f)$,

$$h_i(f) \quad = \quad \sum_{k=1}^{+\infty} \int_{-\infty}^{+\infty} \delta((\phi - k f_h[i]) - f)\, W(\phi)\, d\phi \quad (1)$$

The triangle window function only very roughly approximates the Hann window frequency response. At its base, we assign it the width Δf_t; its peak is assigned a magnitude of unity.

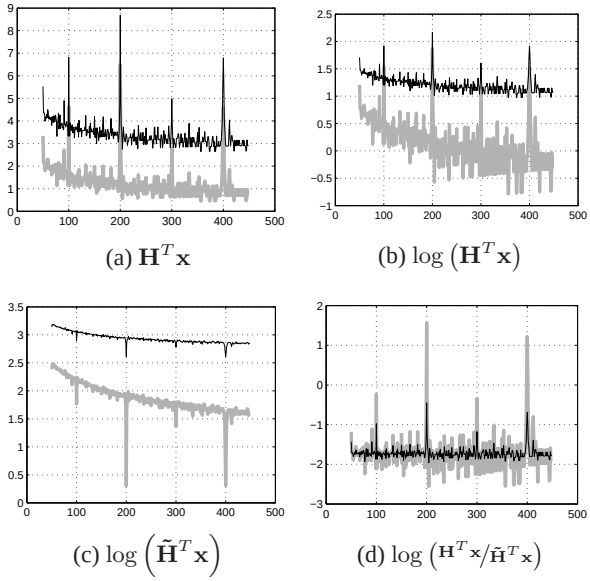
We then Riemann sample each continuous comb filter $h_i(f)$ at the sampling quefrency imposed by our FFT frequency bin centers $f_c[j]$, as shown in Figure 1. (For the first twelve filters, corresponding to candidate fundamental frequencies below 63 Hz, this leads to aliasing since the frequency bin centers are $\Delta f_c = 31.25$ Hz apart; in this first work on this topic, we ignore that concern entirely.) The resulting discrete filters $\mathbf{h}_i[j]$ form the columns of our transform $\mathbf{H}$; two examples are shown in Figure 2. We note that such comb filters rarely appear harmonically spaced, due to discrete sampling.



(a) $F_0 = 100$ Hz (b) $F_0 = 50$ Hz

Figure 2: Discrete-frequency comb filters $\mathbf{h}_i$, in gray, produced by Riemann sampling of continuous-frequency filters $h_i(f)$, in black, with triangular harmonics of base width $\Delta f_t \longrightarrow 0$. The example in (b) violates the Nyquist quefrency criterion; that in (a) does not.

As is performed at the output of the Mel-frequency filterbank prior to decorrelation, we take the (natural) logarithm of the transformed vector, $\mathbf{y} = \log(\mathbf{H}^T\mathbf{x})$. The resulting space of $\mathbf{y}$ is shown for an ideal spectrum $\mathbf{x}$ whose fundamental frequency is 200 Hz, in Figure 3. The spectrum is simply one of the columns of $\mathbf{H}$, with added white noise; the gray line shows $\mathbf{y}$ when $\mathbf{x}$ has a signal-to-noise ratio (SNR) of 10, while that in black is for a SNR of 1. Panel (a) is quite similar to the $\mathrm{SNR} = \infty$ spectra in [9].



(a) $\mathbf{H}^T\mathbf{x}$ (b) $\log\left(\mathbf{H}^T\mathbf{x}\right)$

(c) $\log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right)$ (d) $\log\left({}^{\mathbf{H}^T\mathbf{x}}/_{\tilde{\mathbf{H}}^T\mathbf{x}}\right)$

Figure 3: Various intermediate representations leading to Equation 3. The idealized spectrum $\mathbf{x}$ was formed by adding a column of $\mathbf{H}$, that corresponding to $F_0 = 200$ Hz, to white noise. Gray denotes a SNR of 10, black that of 1.

3.3. Within-Frame Normalization

To reduce the potential effects of noise in our transformed space $\mathbf{y} \in \mathbb{R}^{(N_h+1)}$, we propose to normalize the energy found at each candidate harmonic frequency and its harmonics by the energy found elsewhere. We form the complement transform $\tilde{\mathbf{H}}$, also $\in \mathbb{R}^{(N_f+1)\times(N_h+1)}$, by computing the element-wise additive complement of each entry of $\mathbf{H}$,

$$\tilde{\mathbf{H}}[i][j] \quad = \quad 1 - \mathbf{H}[i][j] \quad . \quad (2)$$

Then the normalized feature vector, used in the experiments in this paper, is given by

$$\mathbf{y} \quad = \quad \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right) \quad . \quad (3)$$

The resulting space of the normalized $\mathbf{y}$ is shown in panel (d) of Figure 3. As can be seen, the proposed normalization eliminates the differences in vertical offset observed for different SNR values, leading only to a difference of magnitude (which also appears to encode SNR).

3.4. Feature Decorrelation

Figure 3 demonstrates that the elements of $\mathbf{y}$ are correlated. In particular, any two elements corresponding to p/q mutliples of the same candidate fundamental frequency will show significant correlation during voicing. We note that this is true not

only of the transformed-domain of $\mathbf{y}$, but of all domains, to our knowledge, in which pitch trackers pick their peaks. It is these correlations which are responsible for octave errors.

For classification tasks, the existence of harmonically-distributed energy at higher frequencies presents a rather important opportunity, provided it is not discarded or is discarded only after consideration (via decorrelation). There exists an abundance of algorithms for decorrelating data, and in our experimental section we explore two of the most laconic: principal component analysis (PCA) and linear discriminant analysis (LDA). Both transform types $\mathcal{F}^{-1} \in \{\mathcal{F}_{PCA}^{-1}, \mathcal{F}_{LDA}^{-1}\}$ provide the opportunity to eliminate candidate fundamental frequencies which are irrelevant to the population of test set speakers. LDA is additionally likely to prefer those candidate fundamental frequencies which actually discriminate among them. Application of such a transform leads to our proposed *harmonic structure cepstral coefficients (HSCCs)*

$$HSCC \quad = \quad \mathcal{F}^{-1} \log\left(\mathbf{H}^T \mathbf{x}\right) - \mathcal{F}^{-1} \log\left(\tilde{\mathbf{H}}^T \mathbf{x}\right) \ , \quad (4)$$

where the second term represents the (decorrelated) normalization of the previous section.

3.5. Relationship to Other Comb Filtering Methods

The discrete-frequency comb filters of $\mathbf{H}$ are a direct discrete implementation of the continuous-frequency "harmonic product spectrum" [11],

$$\Sigma(f) \quad \equiv \quad 20 \log_{10} \sum_{k=1}^{K} |X(kf)| \quad , \quad (5)$$

for some finite integer maximum K. Their construction is similar to that of the discrete-frequency "uniform comb" (UC) filters C in [9]. A filterbank $\mathbf{C}$ of such combs yields a linear-amplitude "PitchPeaks (PP)" representation, which we imply to be given by

$$\mathbf{y}_{PP}^{UC}[i] \quad \equiv \quad \sum_{j=0}^{N_f+1} \mathbf{C}[j][i] \cdot \mathbf{x}[j] \quad . \quad (6)$$

The similarities between our work and [9, 14] end here, because those authors seek to manipulate $\mathbf{y}$ such that $\arg\max(\mathbf{y}_{PP})$ corresponds to the true pitch value. To do so, they first propose the "simple comb" (SC) which limits the number of comb teeth and weights the filter envelope,

$$\mathbf{y}_{PP}^{SC}[i] \quad \equiv \quad \sum_{j=0}^{K_i} \frac{1}{c_i} \mathbf{C}[j][i] \cdot \mathbf{x}[j] \quad , \quad (7)$$

for F_0-dependent tooth limit K_i and scaling function c_i. They then explicitly attempt to suppress the various (p/q) harmonics via the "alternate comb" (AC) filters which contain negative coefficients,

$$\mathbf{y}_{PP}^{AC}[i] \quad \equiv \quad \mathbf{y}_{PP}^{SC}[i] - \sum_{p,q} \xi_{pq} \mathbf{y}_{PP}^{SC}[(p/q)\, i] \quad , \quad (8)$$

for some manually set weights $\{\xi_{pq}\}$.

We could have chosen to also perform these additive and multiplicative operations, and there is the possibility that we are doing so when we infer our $\mathcal{F}_{PCA}^{-1}$ or $\mathcal{F}_{LDA}^{-1}$ transforms. But perhaps not; our task is the discrimination of speakers and not of candidate fundamental frequencies, and the operations may differ significantly.

3.6. Relationship to HNR Computation

For voiced speech frames with true fundamental frequency F_0, exactly one element of $\mathbf{y}$ as given by Equation 3, that at index

$$i^* \quad = \quad \frac{\mathrm{F}_0 - f_h^{min}}{\Delta f_h} \quad , \quad (9)$$

implements a naive variant of a quantity known as the Harmonics-to-Noise Ratio (HNR) [15]. The measure was proposed originally as a tool for diagnosing dysphonic voicing, and many algorithms have been designed for its computation [15, 16, 12, 17]. Estimation, there and in the myriad speech processing systems which have since tried to make use of it, is always conditioned on an available estimate of F_0. We note that HNR is sensitive to (correlated with) both F_0 and jitter.

From the point of view of speech therapists, those values of $\mathbf{y}$ which correspond to candidate fundamental frequencies other than the signal's true F_0 are undefined, as are all values for unvoiced speech.

3.7. Relationship to the FFV Spectrum

Conceptually and functionally, the harmonic structure transform is related to the fundamental frequency variation (FFV) spectrum introduced in [18], and recently applied to speaker recognition in [19]. The FFV algorithm compares the FFT spectrum to a synthetic FFT spectrum, namely the frequency-dilated version of the true FFT spectrum from the preceding speech frame, over a range of logarithmically-spaced dilation factors. This yields a search space; nominally, its $\arg\max$ is the relative change in F_0 expressed in octaves per second.

Here, the HST computation compares the FFT spectrum to another idealized FFT spectrum, namely the discrete-frequency comb filter with known F_0, over a range of linearly-spaced F_0 values. This also yields a search space; nominally, its $\arg\max$ is the absolute F_0 expressed in Hertz.

In both cases, F_0-*independent* dialogue act recognition [20] and F_0-*dependent* speaker recognition, respectively, we opt to model the entire search space rather than its $\arg\max$.

3.8. Relationship to the Mel Filterbank

The filterbank $\mathbf{H}$ has a role which is near-analogous to the well-known Mel filterbank [21], whose matrix formulation we denote as $\mathbf{M}$. MFCC processing consists of computing

$$\mathrm{MFCC} \quad = \quad \mathcal{F}^{-1} \log\left(\mathbf{M}^T \mathbf{x}\right) - \mathrm{MFCC}_{norm} \quad , \quad (10)$$

where $\mathcal{F}^{-1}$ is the data-independent inverse staggered cosine transform and the normalization term is frequently a vector of utterance cepstral means. This equation resembles Equation 4 rather closely.

We note that both filterbanks yield a cepstral coefficient space which is a finite-data approximation of some aspects of

$$\mathrm{CC} \quad = \quad \mathcal{F}^{-1}(\log(\mathbf{x})) \quad , \quad (11)$$

namely the cepstral coefficients obtained without a filterbank. The filterbanks mitigate the need for (near-)infinite data by smearing energy across disparate frequencies, be they related through adjacency (as for $\mathbf{M}$) or harmonicity (as for $\mathbf{H}$). Conceptually, it is important to appreciate that $\mathbf{M}$ and $\mathbf{H}$ mutually destroy each other's input; the spectrum of $\mathbf{M}$ is unrecoverable after $\mathbf{H}$ has been applied, and vice versa[1].

[1] Information pertaining to the vocal source *is* available in the higher-

3.9. Biological Plausibility

It must be admitted that the Mel filterbank [21] owes at least some of its ubiquity to the reputed plausibility of its biological implementation. There are of course many alternatives to the Mel scale [22] on which the filterbank is based, but preferences, for it specifically, are neither undisputed [23] nor germane to this discussion. $\mathbf{M}$ can be any reasonably staggered filterbank whose joint frequency support is contiguous, corresponding to the range of human hearing, and whose individual filters *each have contiguous frequency support*. The physiological evidence for such an $\mathbf{M}$ is: (1) that the basilar membrane implements a one-dimensional tonotopic map ; and (2) that collocated neurons, innervating adjacent hair cells along the organ of Corti, may be reasonably expected to fire together. For completeness, we briefly entertain a similar notion of the plausibility of $\mathbf{H}$, whose individual filters, most notably, *do not* have contiguous frequency support.

We posit that the ability to amplify energies found at integer multiples of a frequency F_0 might also be achieved by collocation, at least in principle, by *rolling* a frequency scale into a circle of circumference F_0. To do so for a range of frequencies $\{F_0\}$ calls for a series of circles. Alternately, it calls for a *spiral*, arranged such that the lowest frequencies map to the smallest radii — not unlike the human cochlea. It is worth mentioning that scientists do not know why the cochlea is coiled [24], short of that it conserves cranial volume or extends frequency range. Almost all studies of human hearing have relied on exclusively one-dimensional models of the basilar membrane, *uncoiled*.

That the cochlea should be additionally a *harmonic frequency analyzer*, rather than merely the housing for a *frequency analyzer* (the basilar membrane) as is widely believed, is most interesting. For if we accept the hypothesis that vocalization and hearing in animal species have co-evolved to benefit *intra*-species communication [25], then, under the above hypothesis, only certain animals should possess a coiled cochlea. Harmonic frequency summation should be of direct value to those species whose members had the ability to control their F_0 *independently* of their vocal tract shape — a function provided for by, most commonly, a glottis. Conversely, we should expect species not possessing a glottis not to have evolved a coiled cochlea. Perhaps serendipitously, both vocal chords and coiled cochleas are near-exclusively unique to mammals; birds, whose vocalizations are considered to be predominantly monophonic whistles, have neither. We close these comments, squarely outside the main scope of the present work, by noting that the average cranial volumes for birds are much smaller than for mammals.

4. HSCC System

The classification system we propose, as elsewhere, accepts audio snippets and produces 1-best hypotheses as to their source. Unlike other systems, our system performs no speech/non-speech segmentation[2]. Every utterance is treated as a contiguous sequence of frames; ideally, we would like the features to be robust towards within-utterance pauses.

4.1. Training

As explained in Section 3.1, audio snippets are first framed; each frame is then windowed, transformed into the discrete-frequency domain, and transformed again via Equation 3. Since we model no inter-frame relationships, all frames are assumed independent, and each is individually associated with a source speaker identity. Given thus labeled TRAINSET and DEVSET audio frames, we proceed as follows.

First, we compute a global decorrelating transform $\mathcal{F}_{PCA}^{-1}$ using TRAINSET, including also the subtraction of a global mean to center the data. This reduces the dimensionality from 400 to, consistently, 397 elements with positive, non-vanishing eigenvalues. Optionally, we compute a second transform, $\mathcal{F}_{LDA}^{-1}$ on top of the first one (and also with its own global mean subtraction), further reducing dimensionality to at most the number of TRAINSET speakers less one.

Second, we optimize the number N_D of decorrelated dimensions by minimizing the speaker classification error on DEVSET. Using TRAINSET data, we train for each speaker a multidimensional Gaussian mixture model (GMM) with a single diagonal-covariance Gaussian. This is a most prosaic model which treats each dimension independently.

Third, we optimize the speaker-independent number N_G of Gaussians in all GMMs by again minimizing the classification error on DEVSET. This time, we retrain the GMMs using the value of N_D optimized above, for a range of N_G values. We make no effort to optimize N_D and N_G jointly, and treat the resulting GMMs as our final models.

Importantly, during this process, no additional data other than that collected from TESTSET speakers in TRAINSET and DEVSET is made use of. In particular, we train no universal background model (UBM), and instead rely on the much simpler to compute maximum likelihood estimates of our models' parameters. While UBMs may ultimately find application for HSCC features, an evaluation of their benefits is beyond the comparative scope of the current work.

4.2. Testing

In testing, snippets from DEVSET or TESTSET are transformed just like those in TRAINSET. Then, every candidate speaker's model is used to estimate the likelihood of all of the snippet's frames. We hypothesize that speaker whose model yields the highest likelihood as the snippet's most likely source.

5. Validation

5.1. Data

Experiments described in the current work use data drawn from the LDC CSR-I (WSJ0) [26] and LDC CSR-II (WSJ1) [27] corpora. Speech snippets consist primarily of read sentences from the Wall Street Journal, but also include some spontaneously produced utterances. We selected them 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, DEVSET, and TESTSET contributions were constructed by including utterances until there were at least 5 minutes of speech data per speaker for training, at least 3 minutes for development, and 3 minutes for

order MFCCs, but of very degraded quality. This is because the Mel filters smear energy across non-harmonically related FFT frequency bins.

[2] In our previous nearfield work [19], on the same data as used here (cf. Section 5.1), we employed an energy-based speech activity detector (SAD) which had been tuned on farfield audio. Since the publication of [19], we have determined that in the nearfield, that SAD algorithm led to very skewed distributions of the number of frames available across speakers. We believe that modifying the SAD algorithm to better match nearfield speech may ultimately mitigate this problem. In the meantime, our much better baseline MFCC results reported here (despite lower model complexity and the absense of a UBM) are achieved without a SAD component.

testing. In total, we identified enough speech data for 95 male speakers and 102 female speakers. Trials initially consisted of 60-second snippets (but we also formed alternative durations of 30 seconds and 10 seconds, as explained below).

Although the data in TRAINSET, DEVSET, and EVALSET for each speaker is multi-session, the majority of sessions from which DEVSET and EVALSET snippets were drawn are also present in TRAINSET. This may turn out to be important, as session mismatch is known to have serious deleterious effect on the performance of other features used in speaker recognition. Specifically, just over 50% of the speakers, of both genders, are represented by speech snippets drawn from exactly the same 4 sessions in each of TRAINSET, DEVSET, and EVALSET. Approximately 25% of both females and males are represented in all three datasets by data drawn from the same 7 sessions. The remaining 25% of speakers are represented by snippets drawn from 10 to 25 sessions, and in these cases the session overlap among the three datasets is between 75% to 100%.

5.2. Experiments

We demonstrate our optimization of the number N_D of principal components and linear discriminants in Figure 4. The classification accuracies are obtained as described in the second step of Section 4.1, using a single diagonal-covariance Gaussian model per speaker. Because accuracies for the original 60-second snippets in DEVSET were higher than expected, we decided to cut them further into 30-second sub-snippets or, alternately, into 10-second sub-snippets. Performance for all three durations of both the PCA- and LDA-transformed development data are shown.

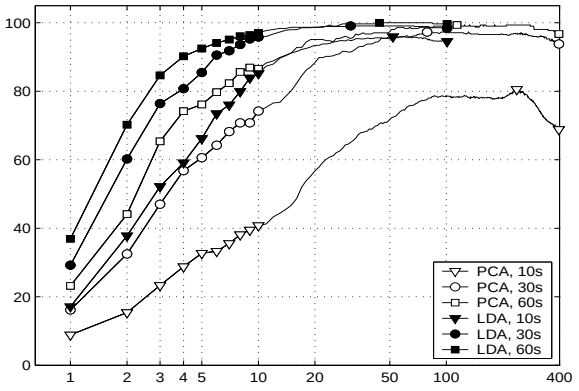


Figure 4: DEVSET accuracy for females, in % along the y-axis, as a function of the number of retained dimensions (along the x-axis). Results for males similar.

As can be seen, accuracies of 99.35%, 97.28%, and 80.56% were achieved for females using PCA, for the three durations of 60 seconds, 30 seconds, and 10 seconds, respectively. As expected, they were higher using LDA: 100%, 99.09%, and 95.94%, respectively. For males (not shown), the PCA results were 100%, 98.46%, 85.42%, respectively, and the LDA results were 100%, 99.81%, 98.67%, respectively. LDA trajectories, almost always, achieved their maxima for smaller numbers of dimensions than did PCA trajectories. We note that all of these results are obtained using one Gaussian per speaker.

We retain only the 10-second condition for the remainder of the experiments in this paper. The number of dimensions, optimized independently for females and males, was also retained throughout. Optimization of the number of Gaussian components per speaker mixture was performed such that the number is identical for females and for males; the best number we found from amongst those that we sampled was $N_G = 256$.

To contrast with HSCC performance, we also trained a standard MFCC system as a baseline. The system retains the first 20 MFCC coefficients, and, for the purposes of comparison, makes no use of a universal background model. On DEVSET, we found that the optimal number of Gaussian components for this system also happened to be 256. Finally, we applied LDA to the MFCC system to produce a variant of the baseline, which more closely matches the design of our HSCC system.

DEVSET and TESTSET accuracies for all three systems, for females and males separately, are provided in Table 1. As the table shows, the accuracy of the proposed HSCC system is higher than that of a baseline MFCC system, with or without LDA. The differences on DEVSET are 1.1%abs for females and 0.3%abs for males; on TESTSET they are 0.6%abs and 1.1%abs for females and males, respectively. Linear equal-weight model-space fusion of the HSCC and MFCC systems yields TESTSET accuracies of 100.00% for females, from 99.87% for HSCCs alone, and 99.87% for males, from 99.65% for HSCCs alone. This translates to two failed trials out of a total of 2922; it also indicates that MFCCs and HSCCs contain complementary information.

Table 1: Classification accuracies (in %) for females (♀) and for males (♂), on both DEVSET and TESTSET. Number of 10 second trials in each split shown in parentheses.

Feature	LDA	DEVSET		TESTSET	
		♀ (1775)	♂ (1660)	♀ (1510)	♂ (1412)
MFCC	no	98.66	99.37	99.27	98.58
MFCC	yes	98.71	99.34	99.27	98.87
HSCC	yes	99.72	99.70	99.87	99.65

Irrespective of the magnitude of the observed differences (but being mindful of their sign), the main consequence of these results is that the HST filterbank provides a representation of the speech signal which is at least as good as the Mel filterbank, for speaker recognition tasks, if not better. But it does so while destroying the spectral envelope. In doing so, it encodes a set of features commonly referred to as (a subset of) prosodic features, which are largely eliminated by the Mel filterbank $\mathbf{M}$. It may be assumed, at this early stage, that HSCCs yield a representation of the glottal source much as MFCCs yield a representation of the vocal tract shape.

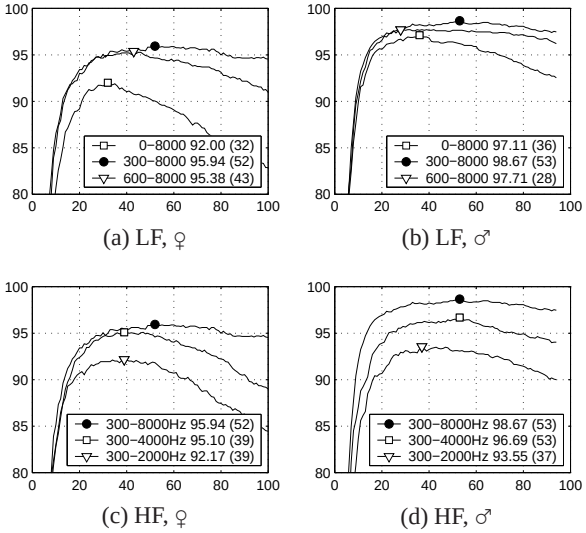
6. Analysis & Discussion

The precise formulation of the proposed transform is governed by many parameters, namely: the sampling frequency f_s; the presence of an optional pre-emphasis filter $1 - 0.97z^{-1}$; the shape and width of the time-signal framing window; the number N_h and spacing Δf_h of the candidate fundamental frequencies; the shape $W(f)$ and assumed width Δf_t of the harmonic model; and alternative normalizations. In this work, these parameters have been set to seemingly reasonable or simply computationally expedient values; no effort has been made to optimize them for speaker recognition or any other task. We leave the majority of such optimization to future work.

The remainder of this section explores the effect on classification accuracy of large perturbations to the frequency magnitude spectrum, prior to computing HSCC features. We have chosen to present only DEVSET numbers, and only using a single diagonal-covariance Gaussian model per speaker, as it has saved us the need to optimize and train more complicated infrastructure. As we have seen, the performance on both DEVSET and TESTSET is very similar, even when we train models with multiple Gaussians and when the number N_G of Gaussians is optimized for DEVSET.

6.1. Ablation of Source-Domain Frequency Range

As a first perturbation, we ablate the frequency range in the magnitude frequency domain prior to computing HSCC features. This consists of simply zeroing out the energy below some low-frequency (LF) cutoff and above some high-frequency (HF) cutoff. The results for LF ablation are shown in panels (a) and (b) of Figure 5. As is evident, the cutoff we have chosen for the systems in our experimental section, of 300 Hz, is the best from among those investigated. Increasing the cutoff to 600 Hz leads to a 4%abs increase in classification error for females, and a 1.5%abs error increase for males. Reducing the cutoff to 0 Hz hurts both genders, but by less. We suspect that the cutoff we chose (actually 296.975 Hz, the first 10 bins in our FFT domain), was overly aggressive and that best performance is to be found for a LF cutoff somewhere between 0 Hz and 300 Hz. A cutoff may not be necessary if signal pre-emphasis is used (as was not done in this work).

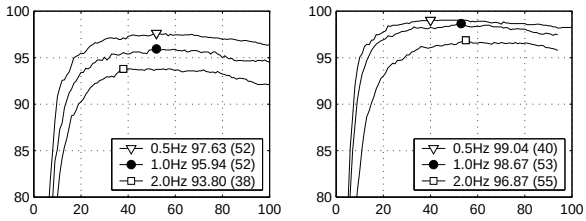


(a) LF, ♀ (b) LF, ♂

(c) HF, ♀ (d) HF, ♂

Figure 5: Effect of ablation of source-domain frequency range on DEVSET accuracy, in % along the y-axis, as a function of the number of LDA coefficients (along the x-axis).

Results for ablating the higher frequencies is shown in panels (c) and (d). Here, ablation always leads to lower accuracies. In particular, for a HF cutoff of 4000 Hz (with the fixed LF cutoff of 300 Hz), corresponding approximately to a standard telephony line, the increase in error is 1%abs and 2%abs for females and males, respectively.



(a) Δf_h, ♀ (b) Δf_h, ♂

Figure 6: Effect of perturbation to frequency resolution on DEVSET accuracy, in % along the y-axis, as a function of the number of LDA coefficients (along the x-axis).

6.2. Perturbation of Transformed-Domain Frequency Resolution

While the perturbation of frequency resolution in the magnitude frequency domain depends only on factors which are in a sense external to the HSCC computation, namely the sampling frequency f_s and the analysis window duration, the frequency resolution in the transformed-domain can be manipulated independently. As mentioned in Section 3.2, we have chosen 400 candidate fundamental frequencies, spaced every $\Delta f_h = 1$ Hz between 50 Hz and 450 Hz. Panels (a) and (b) of Figure 6 show DEVSET performance had we made other choices for the value of Δf_h but kept the same F_0 range.

It appears that the inclusion of more comb filters, spaced more finely in harmonically compressed frequency, yields improvements. They are relatively large for females, of 1.7%abs.

6.3. Ablation of Source-Domain Spectral Envelope

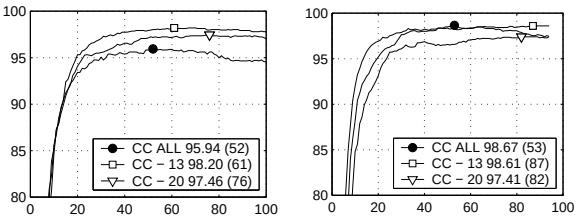
Although we have argued that the HSCC representation is independent of the spectral envelope, as would be directly modeled by the MFCC representation, we haven't provided any empirical evidence. Given that it manipulates the FFT magnitude frequency domain, and that an outcome is that there are *more* dimensions in the transformed-domain ($N_f = 400$) than there were in the source-domain ($N_f + 1 = 257$), one might suspect that decorrelation may recover some envelope information.

To test this hypothesis, we ablate the spectral envelope by truncating the cepstrum, leading to an altered source-domain feature vector $\mathbf{x}'$. First, we transform the log-magnitude frequency domain vector $\log \mathbf{x}$, whose energy is $\mathcal{E}$, via an inverse Fourier transform into the log-magnitude quefrency domain. There, we zero out the first coefficients. We then transform the signal back into the log-magnitude frequency domain, where we undo the log operation, and compute the energy $\mathcal{E}'$. Finally, we normalize the signal such that its energy is again $\mathcal{E}$;

$$\mathbf{x}' \quad = \quad (\mathcal{E}/\mathcal{E}') \cdot e^{\mathcal{F}\left(\mathbf{Z}^T \cdot \mathcal{F}^{-1}(\log \mathbf{x})\right)} \quad , \qquad (12)$$

where $\mathbf{Z} \in \{0,1\}^{2N_f \times 2N_f}$ is a matrix operator which performs the zeroing. The results are shown in Figure 7, for truncation of the first 13 or the first 20 cepstral coefficients, approximating the information available to a "standard" ASR system and to our MFCC speaker recognition baseline, respectively.

We observe that removal of the information contained in the first 13 cepstral coefficients, characterizing the spectral envelope, leads to significant increase in speaker classification accuracy for females, of 2.3%abs. For males there is almost no difference. Further ablation of the next 7 cepstral coefficients hurts



(a) ♀ (b) ♂

Figure 7: Effect of ablation of the spectral envelope on DEVSET nearfield accuracy, in % along the y-axis, as a function of the number of LDA coefficients (along the x-axis). "CC ALL" indicates no ablation, "CC - 13" indicates ablation of the first 13 cepstral coefficients, and "CC - 20" indicates ablation of the first 20 cepstral coefficients.

performance, particularly for males (to 1.3%abs lower than for no perturbation). This provides additional support to claims that higher-order cepstral coefficients (which map roughly to the same-numbered Mel-filterbank cepstral coefficients) contain speaker-specific information. However, those coefficients account for only a fraction of the performance we observe using all the HSCC coefficients, whether ablated in the quefrency domain or not.

7. Conclusions

We have proposed a novel means of modeling instantaneous aspects of prosody, via a harmonic structure transform in magnitude frequency space. Variants of the transformed-space have been repeatedly studied over the past 3 decades; our main contribution consists of resisting the temptation to compute its $\arg\max$, an estimate of pitch. Aside from a normalization, the modeling we have employed is nearly identical to that employed elsewhere in the processing of short-time speech envelope spectra.

The proposed features achieve comparable performance to an MFCC baseline under matched-channel and matched-multisession nearfield conditions. In contrast to prosodic features elsewhere, HSCCs are simple to compute, simple to model, and appear to require neither segmentation nor large quantities of training material. Our analysis suggests that they are robust to various types of gross ablations in frequency space, and that in particular they perform better when the spectral envelope is eliminated.

We believe that the proposed feature space offers a paradigmatic shift in the processing of prosody for speaker recognition, and possibly for other speech processing tasks. Our immediate plans are to examine the features' robustness in the farfield, not only because that condition may most naturally exercise our phylogenetic hypotheses, as well as to derive a compact representation suitable for speaker verification scenarios such as those in the NIST Speaker Recognition Evaluations.

8. References

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

[2] Reynolds, D., Quatieri, T., Dunn, R., "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing, 10(1-3):19–41, 2000.

[3] Doddington, G., "Speaker recognition based on idiolectal differences between speakers", Proc. EUROSPEECH, 2001.

[4] Adami, A., Mihaescu, R., Reynolds, D., and Godfrey, J., "Modeling prosodic dynamics for speaker recognition", Proc. ICASSP, 788–791, 2003.

[5] Weber, F., Manganaro, L., Peskin, B., Shriberg, E., "Using prosodic and lexical information for speaker identification," Proc. ICASSP, 1:141–144, 2002.

[6] Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A., "Modeling prosodic feature sequences for speaker recognition", Speech Communication, 46(3-4):455–472, 2005.

[7] Miller, R. L. and Weibel, E. S., "Measurements of the fundamental period of speech using a delay line", J. Acoustical Society of America, 28:761, 1956.

[8] Moorer, J. A., "The optimum comb method for pitch period analysis of continuous digitized speech", IEEE Trans. Acoustics, Speech, and Signal Proc., 22(5):330–338, 1974.

[9] Liénard, J.-S., Signol, F., and Barras, C., "Speech fundamental frequency estimation using the alternate comb", Proc. INTERSPEECH, Antwerpen, Belgium.

[10] Titze, I. R., *Principles of Voice Production*, Prentice Hall, Englewood Cliffs NJ, USA, 1994.

[11] Shroeder, M. R., "Period histogram and product spectrum: New methods for fundamental-frequency measurement", J. Acoustical Society of America, 43(4):829–834, 1968.

[12] Boersma, P., "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Proc. Institute of Phonetic Sciences, 17:97-110, 1993.

[13] Sun, X., "A pitch determination algorithm based on subharmonic-to-harmonic ratio", Proc. ICSLP, Beijing, China, 2000.

[14] Liénard, J.-S., Barras, C., and Signol, F., "Using sets of combs to control pitch estimation errors", Proc. 155th Meeting Acoustical Society of America, Paris, France, 2008.

[15] Yumoto, E. and Gould, W. J., "Harmonics-to-noise ratio as an index of the degree of hoarseness", J. Acoustical Society of America, 71(6):1544–1550,1982.

[16] de Krom, G., "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals", J. Speech and Hearing Research, 36(2):254-264, 1993.

[17] Qi, Y. and Hillman, R. E., "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals", J. Acoustical Society of America, 102(1):537–543,1997.

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

[19] Laskowski, K. and Jin, Q., "Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum", Proc. ICASSP, Taipei, Taiwan, pp. 4541–4544, 2009.

[20] Laskowski, K., and Shriberg, E., "Comparing the contributions of context and prosody in text-independent dialog act recognition", Proc. ICASSP, Dallas TX, USA, pp. 5374–5377, 2010.

[21] Davis, S. B. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustics, Speech, and Signal Processing, 24(8):357-366, 1980.

[22] Stevens, S. S., Volkmann, J., and Newman, E. B., "A scale for the measurement of the psychological magnitude of pitch", J. Acoustical Society of America, 8(3):185–190, 1937.

[23] Greenwood, D. D., "The Mel Scale's disqualifying bias and a consistency of pitch-difference equisections in 1956 with equal cochlear distances *and* equal frequency ratios", J. Hearing Research, 103(1–2):199-224, 1997.

[24] Cai, H., Manoussaki, D., and Chadwick, R., "Effects of coiling on the micromechanics of the mammalian cochlea", J. Royal Society Interface, 2:341–348, 2005.

[25] Endler, J. A., "Some general comments on the evolution and design of animal communication systems", Philosophical Transactions, 340(1292):215–225, 1993.

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

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