Harmonic Structure Transform for Speaker Recognition

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

[1] KTH Speech, Music and Hearing, Stockholm, Sweden
[2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
kornel@cs.cmu.edu, qjin@cs.cmu.edu

Abstract

We evaluate a new filterbank structure, yielding the harmonic structure cepstral coefficients (HSCCs), on a mismatched-session closed-set speaker classification task. The novelty of the filterbank lies in its averaging of energy at frequencies related by harmonicity rather than by adjacency. Improvements are presented which achieve a 37%rel reduction in error rate under these conditions. The improved features are combined with a similar Mel-frequency cepstral coefficient (MFCC) system to yield error rate reductions of 32%rel, suggesting that HSCCs offer information which is complimentary to that available to today's MFCC-based systems.

Index Terms: speaker recognition, signal processing, harmonic strucure, spectral analysis

1. Introduction

Speaker recognition is quickly becoming a key technology in today's society. Recent years have seen a surge in interest, particularly in feature modeling [1, 2]. Modeling, it is argued, is likely to continue to attract attention. Meanwhile, the spectral features being modeled are predominantly ones that have long been in use; counter-intuitively, their development had been motivated by a need for speaker-independent representations of speech. The most cited example are the Mel-frequency cepstral coefficients (MFCCs).

A key attraction of MFCCs is the simplicity of their computation; prior to decorrelation, the short-time Fourier spectrum is merely passed through a filterbank. The latter averages energy over *contiguous* intervals of frequency. Several alternatives to the filterbank structure have been proposed for speaker recognition [3, 4, 5], with promising results.

In the current work, we explore an altogether different design, in which the filters average energy over frequencies related by harmonicity rather than by adjacency. This renders the frequency support of each filter *non-contiguous*, and destroys spectral envelope shape information. Our starting point is the only direct application of this filterbank to the discrimination of speakers (rather than of fundamental frequencies) that we know of, namely [6]. The closed-set speaker classification experiments presented there relied on 16-kHz close-talk-microphone speech, much of it not spontaneous, under matched channel and matched session conditions. For 10-second trials, the representation — referred to as the *harmonic structure cepstral coefficients* (HSCCs) — yielded accuracies at least as good as those obtained with a comparable MFCC system.

The current paper first evaluates HSCCs on a more spontaneous data set, with session mismatch, otherwise retaining the closed-set classification paradigm for comparison. It is shown that accuracy is only 68% under these conditions, 17%abs lower than that of the MFCC system. Second, we propose and evaluate several novel modifications to the filterbank used to compute HSCCs. The new formulation yields a classification error reduction of 11.7%abs, or 37%rel. Finally, we show that linearly combining the improved HSCC log-likelihoods with MFCC log-likelihoods, thereby merging harmonic structure and spectral envelope information, reduces the error rates achieved by the MFCC system alone by 32%rel.

2. Data

Experiments are conducted on the MIXER5 Corpus [7], a collection of face-to-face interviews containing different types of speech. All participants took part on three days, each of which involved two 30-minute interviews separated by a $\geq$30-minute break. The audio was sampled at 16 kHz in 16-bit quality.

Given the amount of speech available, we selected 66 (39 female and 27 male) out of the 70 speakers recorded at the LDC in Philadelphia, PA. All of our trials come from channel 2, recorded with a Shure MX185 Lavalier microphone worn on each participant's clothing (under the chin). Other channels (all far-field) were not considered in the current work.

In selecting sessions, we located those from which we could obtain the maximum amount of speech for the least-productive of our 66 speakers. These turned out to be sessions 2, 3, and 5 (the latter was least like the other two in that it did not include "sentence reading"). We identified at least 500 seconds of speech from each participant in all three of them. 90 seconds of training material were drawn from session 2 for TRAINSET; 10-second trials were drawn from session 5 for DEVSET and from session 3 for EVALSET. The total number of trials in the two test sets is 3041 and 3045, respectively.

3. Baseline Systems

3.1. Feature Processing

The audio of each trial is framed, without pre-emphasis, into 32-ms Hann windows every 8 ms. A 512-point FFT then yields a 257-point magnitude frequency signal $\mathbf{x}$. We multiply $\mathbf{x}$ by a filterbank matrix $\mathbf{H}$ [6], which consists of $N_h = 400$ rows, each corresponding to one candidate fundamental frequency

$$f_h[i] \quad = \quad f_h^{min} + \frac{i}{N_h}\left(f_h^{max} - f_h^{min}\right) , \qquad (1)$$

with $0 \leq i < N$. $f_h^{min} = 50$ Hz and $f_h^{max} = 450$ Hz; the constant spacing between any $f_h[i]$ and $f_h[i+1]$ is $\Delta f_h = \left(f_h^{max} - f_h^{min}\right)/N_h = 1$ Hz.

The rows of $\mathbf{H}$ are discrete comb filters with 257 entries. Each row is constructed by Riemann sampling a continuous-frequency comb, consisting of a sequence of triangular-shaped

teeth whose apices fall on multiples of the ith filter's $f_h[i]$. Each tooth, of each filter, has a base with a width of $\Delta f_t = 32.25$ Hz and unity area.

$\mathbf{H}$ also has a unity-additive complement $\tilde{\mathbf{H}} = 1 - \mathbf{H}$. HSCCs are obtained by applying a linear discriminant (LDA) transform to the vectors $\mathbf{y}$, where

$$\mathbf{y} = \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right) . \quad (2)$$

As in [6], signal energy below 306.375 Hz is zeroed prior to multiplication by $\mathbf{H}$ or $\tilde{\mathbf{H}}$. Each coefficient of $\mathbf{y}$ represents the log-ratio of energy found at integer multiplies of a putative fundamental frequency to that found everywhere else.

3.2. Modeling and Classification

Speakers are modeled with Gaussian mixture models (GMMs), estimated via maximum likelihood using each speaker's TRAINSET data only. A universal background model (UBM) is not used, and its applicability to the HSCC representation remains a subject for future exploration. We rely on the chunk segmentation in the MIXER5 Corpus, and employ no additional speech activity detector. This means that very short pauses may be included in training and testing.

The classification system is as described in our previous work [6]. The sequence of test feature vectors is scored by each speaker's model; the system hypothesizes that speaker whose model best accounts for the observed vector sequence. Performance is assessed using identification accuracy. The number of frames per trial, under our conditions, is 1250.

We optimize the speaker-independent number N_D of decorrelated dimensions by maximizing accuracy on DEVSET, using a single diagonal-covariance Gaussian classifier. The speaker-independent number N_G of Gaussians in all GMMs is then identified with N_D fixed, again by maximizing DEVSET accuracy[1]. The baseline system, denoted `base` in Table 1, achieves an accuracy of 59.8% on DEVSET, and 68.1% for EVALSET. This represents a lack of robustness to session mismatch, since we observe accuracies of 100% when disjoint training and test data are drawn from the same session.

3.3. Contrastive F_0 and MFCC Features

The supremum of the transformed-domain spectrum $\mathbf{y}$, or of others similarly constructed [8], corresponds to the fundamental frequency (F_0) of the signal in $\mathbf{x}$. The HSCC vector is believed to contain information beyond F_0 [6]. To test this hypothesis, we compare HSCC performance to that achieved by modeling F_0 only. We obtain estimates using the Snack Sound Toolkit [9], and model voiced frames in the log domain using GMMs. As can be seen in Table 1, classification accuracies achieved with this single feature are much lower than for HSCCs.

We also contrast HSCC performance with that obtained using a more traditional set of features, namely the MFCC vector. We compute these by transforming the first 30 Mel filterbank outputs (MEL) using the staggered inverse cosine transform (DCT), and retaining the first 20 coefficients. We apply utterance-level cepstral mean subtraction (CMS), prior to training or testing. Models are GMMs, as for HSCCs; to make for a fair comparison, we do not use a UBM. The performance of the resulting system is shown in Table 1 as "MEL/DCT".

[1]For all HSCC systems, we found $N_D \in [15, 44]$ and $N_G \in \{256, 512, 1024\}$; details are not shown for all development systems due to a lack of space. However, the parameters of the main systems, benchmarked using EVALSET, are shown in Table 4.

Features	N_D	N_G	DEVSET	EVALSET
HSCC `base`	22	512	59.8	68.1
F0 only	1	8	14.1	16.2
MEL/DCT	20	256	74.4	84.8
	25	512	73.6	82.3
MEL/LDA	20	512	79.4	85.3
	25	512	81.5	87.8

Table 1: Baseline and constrast system classification accuracies on DEVSET and EVALSET.

To make comparison with HSCCs more fair, we also replace the data-independent DCT transform with a data-dependent global LDA transform, such as used in computing HSCCs. Table 1 shows that with the same number of coefficients (ordered by eigenvalue), the "MEL/LDA" achieves error rates which are 5.0%abs and 0.5%abs better that "MEL/DCT" on DEVSET and EVALSET, respectively. Selecting the number of coefficients by maximizing the accuracy of a 1-Gaussian classifier on DEVSET, yielding $N_D = 25$, improves on these numbers by 2.1%abs and 2.5%abs, respectively.

4. Development Experiments

We describe several experiment suites treating the optimization of linear candidate fundamental frequency spacing in the filterbank, the elimination of quefrency aliasing, the optimization of logarithmic candidate fundamental frequency spacing in the filterbank, and score fusion with Mel-based systems.

4.1. Linearly Spaced Filterbank Filters

First, we explore the accuracy of the HSCC system while varying the number N_h of filters in the filterbank $\mathbf{H}$, as well as f_h^{min} and f_h^{max} (cf. Equation 1). Table 2 shows the baseline system in bold as `Lin1e` ("`Lin`" refers to uniform spacing of f_0 candidates in the interval $\left[f_h^{min}, f_h^{max}\right]$, with the integer following "`Lin`" indicating the number of sub-intervals). What is clear from this table is that holding $N_h = 400$ and $f_h^{min} = 50$ Hz constant while repeatedly *lowering* f_h^{max} by 50% (e.g., `Lin1i`, `Lin1m`, and `Lin1q`) leads to better performance, until Δf_h reaches a density of 0.25 Hz per filter at $f_h^{max} = 150$ Hz. Also, keeping f_h^{max} fixed but moving f_h^{min} to $1/2\left(f_h^{max} - f_h^{max}\right)$ (e.g., `Lin1g` versus `Lin1e`, `Lin1k` versus `Lin1i`, `Lin1o` versus `Lin1m`) always leads to worse-performing systems. It appears that, given any interval, its lower-order half achieves higher accuracies than its higher-order half.

Informed by this observation, we constructed filterbanks consisting of *two sets* of filters, with different f_h densities. The second panel in Table 2 shows the results. As an example, `Lin2c` is a filterbank in which $N_h = 400$ filters span the sub-range $\left[f_h^{min} = 50 \text{ Hz}, f_h^{max} = 250 \text{ Hz}\right)$ and another $N_h = 400$ filters span the equal-size sub-range $\left[f_h^{min} = 250 \text{ Hz}, f_h^{max} = 450 \text{ Hz}\right)$. This two-sub-range filterbank is seen to perform worse than `Lin2d`, which is identical except that the second sub-range is spanned by only $N_h = 200$ filters. Both pairs `Lin2f` versus `Lin2e`, and `Lin2h` versus `Lin2g`, exhibit the same trend.

Extending this argument to 3-, 4- and 5- sub-range filterbanks, we found that, using a 1-Gaussian classifier, `Lin4a` outperformed the other alternatives. This suggests that the preferred filterbank form is one in which the f_h frequencies are

Feat.	Number of filters, in range (100 | 150 | 250 | 450 | 850)	A, w/ N_G $=1$	A, w/ N_G >1
Lin1a	400	34.4	*56.7*
Lin1b	200	26.8	*51.6*
Lin1c	400	27.9	*56.7*
Lin1d	200	27.2	*49.6*
Lin1e	400	**38.0**	**59.8**
Lin1f	200	26.5	*48.2*
Lin1g	400	28.5	*56.7*
Lin1h	200	28.9	*48.5*
Lin1i	400	42.2	63.9
Lin1j	200	28.4	*52.5*
Lin1k	400	30.3	*60.4*
Lin1l	200	37.1	*59.4*
Lin1m	400	42.4	**67.7**
Lin1n	200	26.8	*54.5*
Lin1o	400	33.3	*64.7*
Lin1p	200	41.6	*65.0*
Lin1q	400	42.0	66.5
Lin2a	400 400	41.6	*64.6*
Lin2b	400 200	40.2	*62.5*
Lin2c	400 400	42.3	*65.4*
Lin2d	400 200	42.8	*66.1*
Lin2e	400 400	43.1	66.5
Lin2f	400 200	43.5	66.8
Lin2g	400 400	42.0	67.2
Lin2h	400 200	42.3	67.1
Lin3a	400 200 200	44.0	*65.1*
Lin3b	400 200 200	44.1	66.6
Lin3c	400 200 200	43.4	*66.3*
Lin4a	400 200 200 200	**45.4**	66.3
Lin4b	400 200 200 200	44.1	*66.5*
Lin5a	400 200 200 200 200	45.3	*65.5*

Table 2: DEVSET accuracies (A), in %, for different linearly-spaced-F0 filterbanks, achieved first with a single-Gaussian classifier ($N_G = 1$) to select N_D and then $N_G > 1$ with N_D fixed (cf. Subsection 3.2). Per row, the number N_h of filters is shown in white, spanning the frequency support in Hz indicated on line 2; the left-most edge is at 50 Hz. Accuracies in italics are linear estimates based on limited N_G optimization efforts.

logarithmically spaced, at least in a piece-wise fashion. Multi-Gaussian experiments ($N_G > 1$), however, indicate that the single-sub-range filterbank `Lin1m` may be the better option.

4.2. Eliminating Quefrency Aliasing

As demonstrated in [6], constructing comb filters with $f_h < 62.5$ Hz leads to what we have called *quefrency aliasing*; this is because, for 16-kHz signals and 512-point FFTs, frequency bin centers are 31.25 Hz apart. To avoid this phenomenon, f_h^{min} is henceforth moved to 62.5 Hz. For the `Lin1m` system of the previous subsection, this yields DEVSET accuracies of 39.9% and 63.1% for single-Gaussian and multi-Gaussian models, respectively (versus 42.4% and 67.7% in Table 2). Accuracies for the `Lin4a` system are reduced to to 43.4% and 65.4%, respectively (versus 45.5% and 66.3% in Table 2); we denote the modification with "`+CUT`" in Table 3. It is not known at this time why the $f_h < 62.5$ Hz filters help; we aim to explore this issue in subsequent work. For the purposes of the current paper, elimination of quefrency aliasing renders the `Lin4a` filterbank more competitive than the `Lin1m` filterbank (65.4% vs 63.1%).

4.3. Logarithmically Spaced Filterbank Filters

Given the evidence in Subsection 4.1, we construct filterbanks whose inter-f_h spacing is continuously logarithmic,

$$f_h[i] \quad = \quad f_h^{min} \left(\frac{f_h^{max}}{f_h^{min}} \right)^{i/N_h} . \tag{3}$$

We express density at a reference frequency of 100 Hz,

$$\Delta f_h \quad \equiv \quad \left. \frac{df_h}{di} \right|_{f_h = 100\,\mathrm{Hz}} \quad = \quad \frac{100\,\mathrm{Hz}}{N_h} \cdot \log_e \left(\frac{f_h^{max}}{f_h^{min}} \right) \tag{4}$$

since it varies with frequency. Our initial logarithmic system, denoted `Log1`, has the same f_h^{min}, f_h^{max}, and N_h as `Lin4a`. Its performance is shown in Table 3; as can be seen, it is only slightly worse than `Lin4a` (also shown). We then move f_h^{min} to 62.5 Hz, as in Subsection 4.2, to avoid quefrency aliasing, yielding filterbank `Log1+CUT`. The drop in performance from `Log1` is similar to that observed in Subsection 4.2.

Feat.	Δf_h (Hz)	f_h^{min} (Hz)	f_h^{max} (Hz)	N_h	A, w/ N_G $=1$	A, w/ N_G >1
base	1.00	50.0	450	400	38.0	59.8
Lin4a	—	50.0	850	1000	45.4	66.3
+CUT	—	62.5	850	950	43.4	65.4
Log1	0.28	50.0	850	1000	45.3	66.0
+CUT	0.28	62.5	850	921	43.4	64.7
Log2	0.28	62.5	4000	1468	50.6	70.2
Log3	0.37	62.5	4000	1129	**51.2**	**70.9**

Table 3: DEVSET accuracies (A), in %, for different filterbanks, achieved first with a single-Gaussian classifier ($N_G = 1$) to select N_D and then $N_G > 1$ with N_D fixed (cf. Subsection 3.2).

Next, we explore the effect of extending f_h^{max}, while holding $\Delta f_h = 0.28$ Hz constant. The best cutoff, at $f_h^{max} = 4000$ Hz, is found for the system denoted `Log2` in the table. It significantly improves accuracies, by 7.2%abs and 5.5%abs for the $N_G = 1$ and $N_G > 1$ classifiers, respectively.

Lastly, varying the density Δf_h while holding $f_h^{min} = 62.5$ Hz and $f_h^{max} = 4000$ Hz constant identifies a system with $\Delta f_h = 0.37$ Hz as optimal; it is denoted `Log3`. The observed improvement in accuracy is 0.6–0.7%abs, despite fewer filters.

4.4. Score-Level Fusion with Mel-based Systems

Finally, we combine the `Log3` HSCC log-likelihoods with those provided by our contrastive Mel-based system, via linear interpolation $\log P = (1-\alpha) \log P_{MFCC} + \alpha \log P_{HSCC}$; a weight of $\alpha = 0$ corresponds to the Mel-based system alone. α is selected by maximizing accuracy on DEVSET.

As is clear from the figure, interpolation with `Log3` HSCC log-likelihoods helps. 20-dimensional Mel-based vectors yield better-performing interpolated systems than 25-dimensional

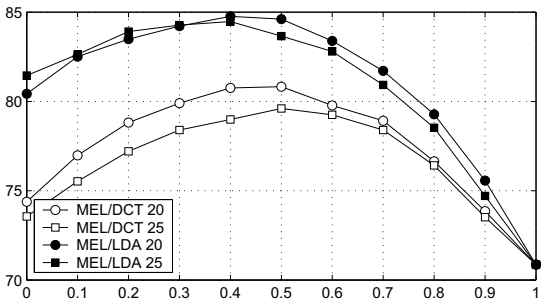

Figure 1: DEVSET accuracies (along y-axis) achieved by linearly interpolating a Mel-system log-likelihood (on the left, $\alpha = 0$) with one produced by the `Log3` HSCC system (on the right, $\alpha = 1$) using weight α (along x-axis).

vectors. Optimal weighting in all cases was found to be 0.5 or 0.4 (in favor of Mel-based systems). The reduction of error is 6.4%abs for the 20-dimension "MEL/DCT" system, and 4.4%abs for the 20-dimension "MEL/LDA" system; these correspond to 25%rel and 22%rel, respectively. The results suggest that MFCCs and HSCCs contain complementary information.

5. Generalization to Unseen Data

We now apply selected manipulations to EVALSET, as shown in Table 4. The major trends observed for DEVSET appear to generalize well. Increasing the f_h range and density, with either a piecewise logarithmic (`Lin4a`) or a continuously logarithmic (`Log1`) structure, results in a 4.4-4.6%abs improvement in classification accuracy. Surprisingly, the subsequent reduction due to our attempts to avoid quefrency aliasing yields only a 0.6%abs drop for `Log1` and no difference for `Lin4a`. The increase in the number of logarithmically inter-spaced filters for candidate fundamental frequencies up to 4 kHz (`Log2`) yields an improvement of 6.0%abs. An additional 1.7%abs is obtained by reducing inter-filter density.

Features	$N_G = 1$		$N_G > 1$	
	N_D	A	N_G	A
HSCC `base`	22	44.3	512	68.1
`Lin4a`	21	50.8	512	72.5
`+CUT`	22	49.6	512	72.5
`Log1`	23	50.8	512	72.7
`+CUT`	29	49.6	512	72.1
`Log2`	27	57.8	256	78.1
`Log3`	27	59.2	512	**79.8**
MEL/DCT	—	—	—	84.8
MEL/DCT $\oplus$ `Log3`	—	—	—	**89.0**
MEL/LDA	—	—	—	84.3
MEL/LDA $\oplus$ `Log3`	—	—	—	**89.3**

Table 4: EVALSET accuracies (A), in %, for different filterbanks designed using DEVSET (cf. Tables 2 and 3 and Figure 1). Also shown are N_D and N_G, optimized using DEVSET. "$\oplus$" denotes score-level fusion.

We observe improvements over the Mel-based systems when interpolating with `Log3` log-likelihoods, of a magnitude approximately equalling that seen for DEVSET. Over "MEL/DCT", the reduction of error is 4.2%abs or 28%rel. Over "MEL/LDA", the reduction is 5.0%abs or 32%rel.

6. Conclusions & Future Work

Session mismatch in nearfield same-microphone speech recordings appears to significantly degrade the performance of the HSCC features proposed in [6]. The current work has presented two improvements to the HSCC representation to address this problem, which yield a reduction in error rate of 11.7%abs or 37%rel for our unseen EVALSET session. The improvements consist of the replacement of constant spacing between the modeled candidate fundamental frequencies by a logarithmic mapping, and an increase in the upper bound for those frequencies. The final HSCC system — whose performance is 5%abs lower than that of a comparable MFCC system — combines well with the latter to achieve relative error reductions of 32% on unseen data, over MFCC performance alone.

We intend to pursue the current work, with the aim of reducing the size of the still very large HSCC vector, and of identifying an appropriate data-independent decorrelating transform. Our final goal is to make the representation sufficiently compact to easily apply the feature modeling techniques (e.g. universal background models) used so successfully with other spectral representations. We anticipate that these measures will enable the assessment of HSCCs on larger problems, such as those in the NIST Speaker Recogntion Evaluations.

7. Acknowledgments

The work was supported in part by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E *Prosody in conversation*.

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010, doi:10.1016/j.specom.2009.08.009.

[2] E. Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I*. 2007, vol. 4343 of *Lecture Notes in Computer Science*, pp. 241–259, Springer, doi:10.1007/978-3-540-74200-5_14.

[3] D. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994, doi:10.1109/89.326623.

[4] T. Kinnunen, "Designing a speaker-discriminative adaptive filter bank for speaker recognition," in *Proc. ICSLP*, Denver CO, USA, 2002, pp. 2325–2328.

[5] H. Lei and E. Lopez-Gonzalo, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 2323–2326.

[6] K. Laskowski and Q. Jin, "Modeling prosody for speaker recognition: Why estimating pitch may be a red herring," in *Proc. ODYSSEY*, Brno, Czech Republic, 2010.

[7] L. Brandschain, C. Cieri, D. Graff, A. Neely, and K. Walker, "Speaker recognition: Building the Mixer 4 and 5 Corpora," in *Proc. LREC*, Marrakech, Morocco, 2008, pp. 3551–3554.

[8] J.-S. Liénard, C. Barras, and F. Signol, "Using sets of combs to control pitch estimation errors," in *Proc. ACOUSTICS*, Paris, France, 2008.

[9] K. Sjölander, "The Snack Sound Toolkit," [software], `http://www.speech.kth.se/snack/`.