Mel, Linear, and Antimel Frequency Cepstral Coefficients in Broad Phonetic Regions for Telephone Speaker Recognition

Howard Lei[1], *Eduardo Lopez*[1,2]

[1]The International Computer Science Institute, Berkeley, CA
[2]Dep. of Signals, Systems and Radiocomm., Universidad Politecnica Madrid, Spain

hlei@icsi.berkeley.edu, eduardo.lopez@upm.es

Abstract

We've examined the speaker discriminative power of mel-, antimel- and linear-frequency cepstral coefficients (MFCCs, a-MFCCs and LFCCs) in the nasal, vowel, and non-nasal consonant speech regions. Our inspiration came from the work of Lu and Dang in 2007, who showed that filterbank energies at some frequencies mainly outside the telephone bandwidth possess more speaker discriminative power due to physiological characteristics of speakers, and derived a set of cepstral coefficients that outperformed MFCCs in non-telephone speech. Using telephone speech, we've discovered that LFCCs gave 21.5% and 15.0% relative EER improvements over MFCCs in nasal and non-nasal consonant regions, agreeing with our filterbank energy f-ratio analysis. We've also found that using only the vowel region with MFCCs gives a 9.1% relative improvement over using all speech. Last, we've shown that a-MFCCs are valuable in combination, contributing to a system with 17.3% relative improvement over our baseline.

Index Terms: Speaker recognition, MFCCs, LFCCs, a-MFCCs, filterbank analysis

1. Introduction

Mel-frequency cepstral coefficients (MFCCs) have been widely used as features for speaker recognition, primarily because they've been empirically determined to work well for speaker recognition after being developed for speech recognition [1]. The coefficients rely on a mel-frequency spacing of filterbank energies, which mimics the frequency response of the human ear. [2] has shown using a set of 35 speakers, however, that frequency regions around 300 Hz, 4,500 Hz, and 7,500 Hz hold greater speaker discriminative power than other frequency regions, primarily due to speaker-specific nasal coupling, piriform fossa, and consonants constrictions. Moreover, filterbank energies, when used as features, have been shown to have higher f-ratios for filters near frequency regions with greater speaker discriminative power [2].

Hence, a set of filterbanks different from the mel-spaced filterbanks used in MFCCs from 0 to 8,000 Hz have been proposed by [2], where the filters are more tightly spaced with narrower bandwidths at frequencies with higher speaker discriminative power, and more widely spaced with wider bandwidths at other frequencies. The new filterbank arrangement has more filterbank channels near frequencies of higher speaker discriminative power, and has been found to perform better than MFCCs and LFCCs in a GMM-based speaker recognition system [2].

Normal telephone speech, however, is band-limited to between 300 and 3,400 Hz, and the dominant frequency regions for speaker discrimination are absent. However, we have observed that the f-ratios of a set of linearly-spaced filterbank energies are still higher near 3,400 Hz as opposed to 300 Hz. We've thus experimented with linear- and antimel-frequency filterbank spacings, where filters are more tightly spaced at higher frequencies compared to lower frequencies. Denote the cepstral coefficients using the antimel frequency spacing as a-MFCCs.

In addition to our experimenting with different filterbank frequency spacings, we've also experimented with using only speech data from broad-phonetic regions of speech, such as nasals, vowels, and non-nasal consonants. This is because certain speaker-specific attributes (i.e. nasal coupling) exist only when certain phones are spoken, and because we've observed differences in the filterbank energy f-ratios for different broad-phonetic regions.

This paper is organized as follows: Section 2 describes the database, section 3 describes our f-ratio results and a-MFCC extraction, section 4 describes our speaker recognition system, section 5 describes the experiments and results, section 6 provides a brief discussion, and section 7 provides a summary of our findings.

2. Database

We've used the Switchboard II and Fisher corpora for universal background speaker model training and SRE06 for speaker model training (w/ 1 conversation side) and testing. All corpora consists of telephone conversations between two unfamiliar speakers. A conversation side (roughly 2.5 minutes for non-Fisher and 5 minutes for Fisher) contains speech from one speaker only. $\sim$3,200 conversation sides are used in SRE06, and $\sim$1,550 conversation sides are used for background training. There are $\sim$23,000 total trials with $\sim$1,800 true speaker trials.

We are provided with force-aligned phone ASR decodings for all conversation sides by SRI, obtained via the DECIPHER recognizer [3], from which we are able to determine what data to use for the broad-phonetic regions. The DECIPHER recognizer uses 46 phones (not including the starts and stops), with 4 nasals, 14 vowels, and 28 consonants. Using $\sim$1,100 conversations of Fisher background data, we've determined that, not including the non-speech regions, $\sim$10% of the data contains nasals, $\sim$42% contains vowels, and $\sim$48% contains non-nasal consonants.

3. F-ratio analysis and feature extraction

Similar to what was done in [2], we've extracted a set of 26 uniformly-spaced triangular filterbanks from 300 to 3,400 Hz to determine which frequency regions within the telephone band-

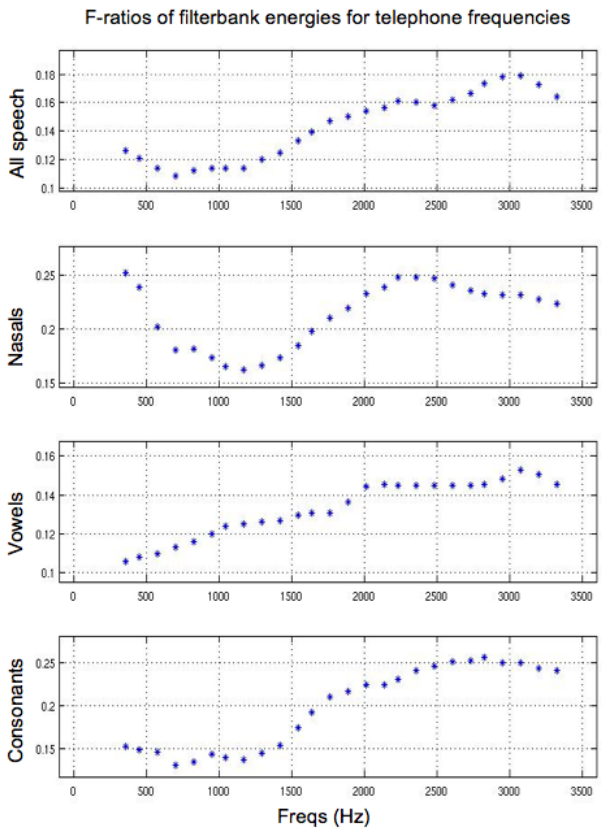

Figure 1: *F-ratio plots of filterbank energies for nasals, non-nasal consonants, vowels, and all speech on gender-balanced SRE08.*

width are useful for speaker discrimination. The bandwidth of each filter is 187.5 Hz, with 62.5 Hz of overlap between adjacent filters. The F-ratio is defined as follows:

$$F-ratio = \frac{\sum_{speaker:s}(\mu_s - \mu)^2}{\sum_{speaker:s}\sum_{i \in s}(x_i - \mu_s)^2} \quad (1)$$

where the summations are across all distinct speakers, μ_s is the average of the filterbank energies within speaker s, μ is the average of filterbank energies across all speakers, and x_i is filterbank energy i in speaker s.

Figure 1 plots the f-ratios of each filterbank energy when the f-ratio is computed using data from all speech, and only the nasals, vowels, and non-nasal consonants on ∼1,000 gender-balanced SRE08 telephone conversation sides. The frequencies of each point correspond to the frequencies of the filterbank centers.

According to the plots, the highly speaker discriminative region near 300 Hz results in a sharp f-ratio increase near 300 Hz for the nasal regions, and a small increase for the non-nasal consonants. The increase in f-ratio near 300 Hz occurs slightly using all speech, probably due to the nasals within the speech. In all cases, the f-ratio increases in the higher frequency regions. The increase suggests that even for telephone speech, where the speaker-dependent regions caused by physiological characteristics are filtered out, using a set of filters more closely spaced in the higher frequency regions for cepstral feature extraction may be desirable.

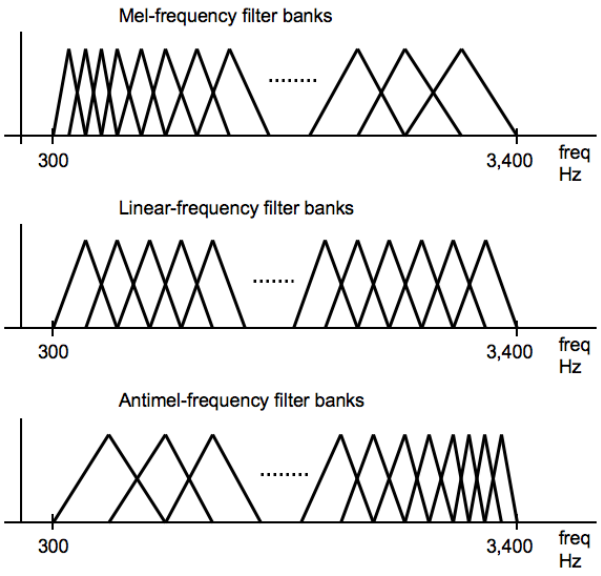

Figure 2: *Illustration of the mel-, linear-, and antimel- frequency filterbanks.*

Note that, as was done in [2], it is possible to extract a set of cepstral coefficients using filterbanks more tightly spaced in both the high and very low frequency regions (according to the f-ratios of each region). However, these features would not be uniformly spaced along any standard frequency axis, and may involve a significant degree of customization for future extraction. The high speaker discriminative regions may also be database-dependent, and may vary depending on the speakers used. Hence, a set of features customized using one dataset may not generalize well to other datasets. We've thus decided to experiment with standard feature sets with filterbanks uniformly spaced along various frequency scales.

We've implemented a-MFCCs via HTK [4], where the mel-spaced filters from 300 to 3,400 Hz are flipped about the middle of the frequency region (1,550 Hz) such that filters at high frequencies are shifted to low frequencies, and vice versa. The a-MFCC filterbank filters can be thought of as being equally spaced on the "antimel-" frequency scale, as opposed to being equally spaced on the mel-frequency scale for the case of MFCCs. LFCCs are also implemented for comparison. Figure 2 illustrates the mel-, linear-, and antimel-frequency filterbanks.

Like the MFCCs and a-MFCCs, the LFCCs also use triangular filterbanks in our experiments. LFCCs may be desirable in that it places more filterbank emphasis on the higher frequency regions with higher filterbank energy f-ratios, without sacrificing too many filterbanks in the lower frequency regions, where the nasals and non-nasal consonants show an increase in filterbank energy f-ratio.

4. Speaker recognition system

To test out the effectiveness of a-MFCCs (and LFCCs as a reference), a 512-mixture GMM-UBM system [5] with factor analysis is used. The 0th through 12th coefficients of either the MFCCs, LFCCs, or a-MFCCs (with 25 ms windows and 10 ms intervals) with deltas and double deltas are used. Mean and variance feature normalization is applied, and the speaker modeling, scoring, and factor analysis are implemented using the

ALIZE toolkit [6]. The simplified factor analysis model with 40 nuisance factors is used to decompose the GMM means into speaker-independent, speaker-dependent, and nuisance factors [7].

5. Experiments and results

We implemented our speaker recognition system on all speech data, and each of the broad phonetic regions. We performed various score-level combination experiments using an MLP (via Lnknet [8]) with 2 hidden nodes and 1 hidden layer, and tried to determine how well a-MFCCs, LFCCs, and MFCCs combine, as well as how the broad phonetic regions combine.

Denote each system using two characters – one for the feature used and one for the broad phonetic region used. a, l and m denote a-MFCCs, LFCCs, and MFCCs respectively, and *c*, *n*, *v* denote non-nasal consonants, vowels, and nasals respectively. In addition, *sysa*, *sysl*, and *sysm* denote the systems using a-MFCCs, LFCCs, and MFCCs respectively on all speech data (excluding the non-speech regions for each conversation side), and *sysm* is taken to be our baseline. Systems using all speech data are regarded as full-systems; others as sub-systems.

All results are obtained on SRE06, and all score-level combinations are performed by randomly creating 40 pairs of training and testing splits of the data (where the number of models and test utterances for each training split is roughly equal to the numbers for the corresponding testing split), and averaging the EERs across all pairs. Table 1 displays the results, along with the portion of data used to implement the system (according to the percentage of nasals, vowels, and non-nasal consonants in our Fisher background data).

Results in table 1 are subdivided into various categories, where category 1 contains results for a-MFCCs, LFCCs, and MFCCs using all speech data, categories 2, 3, and 4 contain standalone results for each of the broad phonetic regions using each of the feature types, category 5 contains results for each of the broad phonetic regions but combining the three types of features for each region, category 6 contains results for each of the features but combining the three types of broad phonetic regions, category 7 contains a combination of all regions and all features, and category 8 contains our best overall result. Note that categories 1, 2, 3, 4, and 6 discriminate between performances of the three features.

MFCCs and LFCCs have produced the best overall results, contrary to the f-ratio plots which suggest that a-MFCCs would be better. In category 1, *sysm* (our baseline system with 5.26% EER) slightly outperforms *sysa* and *sysl*. MFCCs significantly outperformed LFCCs in categories 3, and 6, whereas LFCCs are significantly better in categories 2 and 4. In category 2 (nasal sub-systems) LFCCs show 21.5% and 28.1% relative improvements over MFCCs and a-MFCCs respectively. In category 4 (non-nasal consonants only), LFCCs show 15.0% and 19.3% relative improvements over MFCCs and a-MFCCs respectively.

Also note that vowels are significantly better than other regions of speech, where the average EER of category 3 (vowels standalone) is 5.43%, a 39.7% relative improvement over the nasal average in category 2 (8.99% EER) and a 35.5% relative improvement over the non-nasal consonants average in category 4 (8.40% EER). The better performance of vowels over the nasals may be due to the significant increase (420%) in the amount of vowel data over nasal data in speech.

Interestingly, the sub-system using only the vowel region and MFCCs (4.78% EER) outperforms the three systems in category 1 that use all speech data. We've also discovered that the combination of the 9 sub-systems in category 7 produces a 11.41% relative improvement over our baseline system. We've found the best overall system to be a combination of the *vm*, *vl*, and *nm* sub-systems, along with the *sysa* full-system, producing a 17.3% relative improvement over our baseline system.

Table 1: *Speaker recognition results for systems standalone and in combination using a-MFCCs, LFCCs, and MFCCS as features, and nasals, vowels, non-nasal consonants, and all speech as data. Results obtained on SRE06.*

Category	System combination	EER (%)	Percent data used
1: *All speech*	*sysm*	5.26	100
	sysl	5.35	100
	sysa	5.51	100
	sysm+sysl+sysa	4.96	100
2: *nasals only*	*nm*	9.37	∼10
	nl	7.36	∼10
	na	10.24	∼10
3: *vowels only*	***vm***	**4.78**	∼42
	vl	5.43	∼42
	va	6.05	∼42
4: *non-nasal cons. only*	*cm*	8.68	∼48
	cl	7.38	∼48
	ca	9.14	∼48
5: *feature combination*	*ca+cl+cm*	6.77	∼48
	na+nl+nm	6.40	∼10
	va+vl+vm	5.10	∼42
6: *region combination*	*cm+nm+vm*	4.54	100
	cl+nl+vl	5.39	100
	ca+na+va	5.61	100
7: *overall combination*	*ca+cl+cm+ na+nl+nm+ va+vl+vm*	4.66	100
8: *best EER results*	*vm+vl+nm*	4.41	∼52
	vm+vl+nm+ sysa	**4.35**	100

The standalone nasal sub-systems also use significantly less data than the vowel and consonant sub-systems, which likely contributed to its poorer performance relative to the vowels and consonants. However, the combination of the 3 nasal sub-systems in category 5 slightly outperforms the combination of the 3 non-nasal consonant sub-systems while using ∼21 percent of the data. Table 2 shows the results for the nasals, vowels, and non-nasal consonants for MFCCs when the vowel and non-nasal consonant phone instances are removed such that there're roughly equal portions of data for the three regions. Note that these results are not computing using EER-averaging of the 40 splits, but on the entire set of scores (hence the slight difference in the results for the *vm* sub-system).

According to table 2, the nasal sub-system (9.46% EER) performs 43.0% better than the vowel sub-system, and 53.2% better than the non-nasal consonant sub-system. This agrees

Hence, while the a-MFCCs pale in comparison with MFCCs and LFCCs standalone, it is valuable as a feature in combination with the other features, as it emphasizes different parts of the frequency spectrum compared to the other features. Note that the combination of *vm*, *vl*, and *nm* sub-systems performs better than all other systems and their combinations (not all of which are shown) while using roughly half the total speech data.

Table 2: *Speaker recognition results for sub-systems using MFCCS as features with roughly equal portions of data for the three broad-phonetic regions. Results obtained on SRE06.*

Sub-system	EER (%)
nm	9.46
vm	16.61
cm	20.21

with the preliminary subjective study using 10 speakers and 6 consonants (nasals included) by [9], which suggests that nasals are more effective (but without significance) than other consonants in speaker discrimination [9].

We have shown here, however, that using our GMM-UBM system with factor analysis, nasals significantly outperform vowels as well as non-nasal consonants using MFCCs on roughly the entire SRE06 1-conversation side task. One of the reasons suggested is that the resonators in the nasal tract are highly speaker-dependent, and can not be altered at will [9]. Hence, should constraints on the overall amount of speech data be made, nasals should be amongst the first phones to be included.

6. Discussion

We have shown the value of implementing a separate speaker recognition system on the vowel, nasal, and non-nasal consonant regions, which combined to produce a better result that implementing a single system using all speech data. We have also demonstrated the value of LFCCs using the broad phonetic regions, as the LFCCs performed significantly better than a-MFCCs and MFCCs for the nasal and non-nasal consonant regions.

As suggested in section 3, the desirability of LFCCs is likely due to the fact that LFCCs use more filterbanks compared to MFCCs in the higher frequency regions with the higher filterbank energy f-ratios, without sacrificing too many filterbanks in the low-frequency region, where the non-nasal consonants and especially the nasals show an increase in filterbank energy f-ratios.

The a-MFCCs did not work as well as we expected standalone. While [2] has shown that there is up to a 9-fold increase in filterbank energy f-ratio from the least speaker discriminative to most speaker discriminative frequency regions from 0 to 8,000 Hz, our f-ratio plots show less than a 2-fold increase within the telephone bandwidth. Hence, it is possible that the a-MFCCs sacrifice too many filterbanks at the low-frequency regions where nasals and non-nasal consonants show an increase in filterbank energy f-ratio. However, we have shown that the a-MFCCs improve results in combination with sub-systems using other types of features, contributing to our best overall result.

Future work can perhaps investigate the effectiveness of cepstral coefficients derived from a filterbank spacing with more filters at both the low and the high frequency regions (as done in [2]) in the telephone bandwidth, where increases in f-ratios occur. However, as suggested in section 3, the sets of features we've investigated are standardized and do not require significant customization for future extraction. They are also likely to generalize well to other databases.

7. Conclusions

In this work, we've investigated different broad phonetic regions and standard filterbank spacings for cepstral feature extraction on telephone speech, where highly speaker discriminative frequency regions are filtererd out. We've demonstrated that LFCCs perform significantly better than MFCCs for speaker recognition systems using the nasal and non-nasal consonant broad phonetic regions, agreeing with our plots of filterbank energy f-ratios within the telephone bandwidth. We've also shown that a-MFCCs are useful in overall system combination, contributing to a combined system with 17.3% relative EER improvement over our baseline system.

8. Acknowledgements

The authors wish to thank Andreas Stolcke of SRI for providing speech recognition decodings. This research is funded by NSF grant number 0329258. The second author is at ICSI on sabbatical leave from UPM thanks to the Spanish Science and Innovation Ministry's support of ICSI.

9. References

[1] Reynolds, D.,"Experimental evaluations of Features for Robust Speaker Identification", in IEEE Trans. Speech and Audio Processing, Vol. 2, 1994.

[2] Lu, X., Dang, J.,"Physiological feature extraction for text-independent speaker identification using non-uniform subband processing", in Proc. of ICASSP, 2007.

[3] Stolcke, A., Bratth, H., Butzberger, J., Franco, H., Rao Gadde, V., Plauche, M., Richey, C., Shriberg, E., Sonmez, K., Weng, F. and Zheng, J.,"The SRI March 2000 Hub-5 Conversational Speech Transcription System", in NIST Speech Transcription Workshop, 2000.

[4] HMM Toolkit (HTK), http://htk.eng.cam.ac.uk

[5] Reynolds, D.A., Quatieri, T.F. and Dunn, R.,"Speaker Verification using Adapted Gaussian Mixture Models", in Digital Signal Processing, pp 19–41, 2000.

[6] Bonastre, J.F., Wils, F., Meignier, S.,"ALIZE, a free Toolkit for Speaker Recognition", in Proc. of ICASSP, 2005.

[7] Kenny, P., Dumouchel, P.,"Experiments in speaker verification using factor analysis likelihood ratios", in Proc. of Odyssey, 2004.

[8] Lippmann, R.P., Kukolich, L.C., Singer, E., "LNKnet: Neural Network, Machine Learning, and Statistical Software for Pattern Classification", in Lincoln Laboratory Journal, Vol. 6, pp 249-268, 1993.

[9] Amino, K., Sugawara, T. and Arai, T.,"Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties", in Acoustic Science and Technology, 27(4), 2006.