ON DATA-DERIVED TEMPORAL PROCESSING IN SPEECH FEATURE EXTRACTION

Michael L. Shire, Barry Y. Chen

International Computer Science Institute
University of California at Berkeley
Berkeley, California
{shire,byc}@icsi.berkeley.edu

ABSTRACT

Temporal processing and filtering in speech feature extraction are commonly used to aid in performance and robustness in automatic speech recognition. Among the techniques successfully employed are RASTA filtering, delta calculation, and cepstral mean subtraction. The work here explores the use of temporal filter design using LDA to further enhance performance using a few preprocessing configurations. In addition to RASTA filtering, we apply the filters to modulation-spectral features and cepstra while making sure that the assumptions of LDA are observed. We additionally test the use of filters that have been trained in different reverberation conditions, noting from previous work that the presence of reverberation alters the preferred frequency range of the derived filters. Our tests indicate a consistent advantage in phone classification. Word recognition tests, in contrast, reveal that the LDA filters often do not improve upon the existing filters previously used. They can also be made less effectual by allowing contextual frames to a trained probability estimator.

1. INTRODUCTION

Temporal processing and filtering in speech feature extraction are commonly used to enhance performance and robustness in automatic speech recognition. Techniques such as cepstral mean subtraction (CMS) [1], delta calculation [8], and RASTA filtering [13] are some examples of temporal processing that have been successfully applied. In each of these techniques, the trajectories of spectral or spectrally related values are temporally processed to enhance or preserve the speech carrying modulations or to add speech dynamics information to the recognition system. Sometimes neglected is the implicit temporal processing when training a probability estimator with a context of several adjacent frames of acoustic features. The precise implementation of the explicit temporal filtering techniques was often from insight and repeated empirical testing. Van Vuuren and Hermansky introduced a data-driven technique for direct derivation of temporal basis functions through linear discriminant analysis (LDA) [20]. Their work concerned the derivation of RASTA-style filters from log-critical-band spectral values of a phonetically labeled corpus; phonetic classes were assigned to the log-spectral temporal trajectories. Recent efforts have observed the effects of additive noise and reverberation on these filters [19]. Lieb and Haeb-Umbach have also recently applied the technique to trajectories of Mel-Frequency Cepstral Coefficients (MFCC) [17]. When applied to the feature trajectories, the filter components can be seen as a replacement of the delta calculation.

The original work with temporal LDA typically involved replacing the filters in log-RASTA-PLP with these data-derived filters. Though originally applied to log-spectral values, we find that the results of this method may, with some care, be applied in alternate settings. In this work we make some further observations on the use of LDA for temporal filtering in feature extraction in our recognition system. We apply the temporal LDA technique to a few preprocessing configurations wherein other implementations of temporal filtering were replaced by filters derived through LDA. In addition to log-RASTA-PLP [13], we apply the LDA filters to the modulation-filtered Spectrogram (MFSG) [16, 15] and the original PLP [11] as a postprocess.

2. EXPERIMENTAL SYSTEM

Our experimentation of the use of temporal LDA was within the framework of a hybrid artificial neural network - hidden Markov model (ANN-HMM) automatic speech recognitions system [18, 3]. In this system, a simple three-layer feed-forward multi-layer perceptron (MLP) is discriminatively trained to estimate the posterior probabilities of context-independent mono-phone classes given the acoustic features. In previous systems, particularly those using Gaussian mixture models, LDA was applied to the feature vectors. In other words, discriminative training was applied to the spectrally related dimension. In the ANN-HMM context, LDA applied in this manner is redundant to the non-linear discrimination inherent in the MLP training. LDA however, when applied temporally, adds a layer of discriminative training along the temporal dimension. In effect, we achieve a discriminative training that covers the time-frequency plane.

Temporal filter derivation begins by capturing windows of approximately one seconds worth of spectrally-related trajectories. These trajectories are of the logarithm of filter-bank envelopes or cepstral trajectories. Each window is assigned a phonetic class label that corresponds to the center. The average covariance matrix of the classes S_W and the covariance of the means of the classes S_B are subsequently computed. The eigenvectors of $S_W^{-1} S_B$ that have the corresponding largest eigenvectors are taken as the discriminatively trained filters [2, 20, 7].

In our experiments, temporal filters and recognition tests were performed using separate corpora to promote generality. The filter design used the English portion of the Oregon Graduate Institute (OGI) Multi-Lingual Database [4] that included hand-labeled and segmented phonetic transcriptions. We additionally designed filters with the speech corpus artificially reverberated with two room impulse responses. The first impulse, labeled "light", had the quality of a small office with a reverberation time (T_{60}) of 0.6 seconds and a direct-to-reverberant ratio (DTRR) of -1.9 dB. The second, labeled "heavy", was recorded in a concrete basement hallway having a T_{60} of 2.5 seconds and of DTRR of -8 dB. Subsequent recognition tests were conducted with a

Appears in ICSLP 2000, vol 3, pp 71-4

Beijing China, October 2000

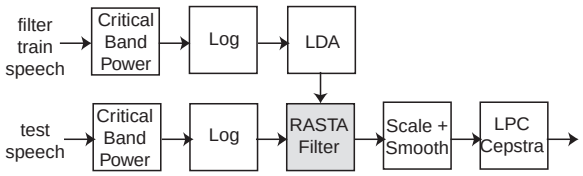

Figure 1: RASTA filters replaced by LDA filters.

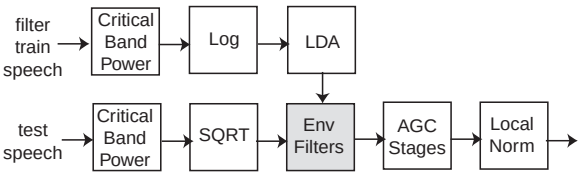

Figure 2: MFSG envelope filters replaced by LDA filters.

MLP environ.		LDA training environment		
train	test	clean	light	heavy
clean	clean	**9.10**	9.00	12.00 -
light	light	18.90	**16.40** +	18.30
heavy	heavy	45.50	42.10 +	**38.90** +

Table 1: WER using LDA filters derived with reverberant data and MLP with single frame of features.

MLP environ.		LDA training environment		
train	test	clean	light	heavy
clean	clean	**5.20**	5.30	7.00 -
light	light	11.10	**10.70**	12.40 -
heavy	heavy	30.70	30.10	**30.30**

Table 2: WER using LDA filters derived with reverberant data and MLP context window of 9 feature frames (10 ms stepping rate).

subset of the OGI Numbers corpus [5] that also had phonetic transcriptions.

3. RASTA-LDA

In previous work we had noted that increasingly severe reverberation resulted in LDA filters that preferred the lower frequency ranges [19]. For example, the first discriminant filter for the clean, light, and heavy reverberation environments used here had upper half-power points at 13 Hz, 9 Hz, and 5 Hz respectively. When the original single-pole RASTA filter was replaced with these LDA filters within RASTA-PLP (figure 1), we observed performance improvements in recognition tests, particularly in cases of reverberation. In these previous experiments, the MLP probability estimator was allowed approximately 100 ms of contextual feature frames. As the bulk of the filter impulse response resided within this range, it is possible for the MLP training to learn and mimic some of the temporal filtering characteristics. To better observe the appropriateness of each filter to the environment with which it was trained, we conducted additional tests comparing the three sets of LDA filters. In the first, we eliminated the contextual frames allowed to an MLP probability estimator with 800 hidden units. Table 1 contains word error rates (WER) of matched training and testing experiments. The "+" ("-") postfixes mark where improvement (degradation) over the clean LDA filters was statistically significant (p=0.05, 4673 words). Boldface signifies matched training, testing, and filter design environments. We observe here that when the probability estimator is not allowed contextual frames, the recognition is best when using the filters trained in the same environment. The only exception appears to be in the clean case, where the light filters perform better by 1% relative, though this is not significant.

Train	Test	MFSG	MFSG-LDA
clean	clean	6.50	7.00
light	light	12.10	12.40
heavy	heavy	31.60	32.80

Table 3: WER comparison of original MFSG and MFSG with LDA-derived filters. MLP trained on acoustic context of 9 frames.

Train	Test	MFSG	MFSG-LDA
clean	clean	76.96	78.40 +
light	light	70.95	72.82 +
heavy	heavy	55.66	57.89 +

Table 4: Frame accuracy comparison of original MFSG and MFSG with LDA-derived filters. MLP trained on acoustic context of 9 frames.

Table 2 contains WER scores when we re-introduced contextual frames to the 800 hidden unit MLP. In this configuration, we no longer see a consistent advantage in using the alternate filter in each training and testing environment; using the light filter seems to produce the best scores in the reverberation tests but the differences are not significant. The heavy filter set, which smoothes the most, appears to be less useful overall since it produces the worst scores in all but the matched testing environment.

4. MFSG-LDA

LDA provided an automatic statistical means of generating RASTA-style filters. These filters demonstrated frequency selectivity in agreement with previous perceptual and empirical data [12, 14, 9]. In an effort to test the general usefulness of these RASTA-style filters, we sought to apply them in other feature processing strategies. The MFSG process was developed by Kingsbury and has demonstrated utility in cases of reverberation [16, 15]. Here we replace the temporal envelope filters included in that processing with RASTA filters derived using LDA, as in figure 2. In contrast to the LDA employment in RASTA-PLP, we do not tap and analyze the trajectories of the amplitude spectra after the square root operation, but rather continue to use the logarithm. The reason for this is that the amplitude spectrum provides a poor domain with which to apply LDA. An assumption of LDA is that the underlying class distributions are normal. Applying a logarithm to the energy envelopes produces distributions that are closer to the normal distribution than the amplitude spectrum. Employing a logarithm is a common technique in normalizing data in statistics as well as speech and has the advantage of being "shape-invariant" to scaling and powers of the raw data. That is, a scaling of the data appears as an offset after a logarithm while a power merely adjusts the spread and domain of the data distribution. We reconcile using filters derived from

Appears in ICSLP 2000, vol 3, pp 71-4

Beijing China, October 2000

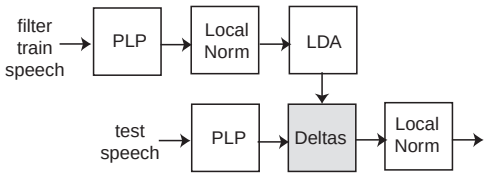

Figure 3: PLP and delta features replaced by LDA filtered features.

Train	Test	PLP+Δs	PLP-LDA	
clean	clean	7.80	8.80	-
light	light	16.20	17.60	-
heavy	heavy	46.10	40.40	+

Table 5: WER comparison of original PLP and PLP with LDA-derived filters. MLP trained on a single frame of acoustic features.

Train	Test	PLP+Δs	PLP-LDA	
clean	clean	70.75	74.65	+
light	light	62.05	67.74	+
heavy	heavy	45.59	53.48	+

Table 6: Frame accuracy comparison of original PLP and PLP with LDA-derived filters. MLP trained a single frame of acoustic features.

logarithm data in the situation where the square root is used by noting that a main effect of memoryless nonlinearities of this type is the creation of harmonics; the fundamental modulations remain. What arguably remains essential is the preservation of modulation rates that are important to discrimination.

The original envelope filters used were 0-8 Hz and 2-16 Hz IIR filters and were arrived at through repeated recognition tests. Tables 3 and 4 show the WER and frame accuracy results when these filters were replaced by the first two LDA components derived from the matched training environment. The frame accuracy is consistently better by between 1% and 4% relative when using the LDA filters. This was also observed in almost all mismatched training and testing conditions not shown here. However, we do not see an improvement in WER; the use of LDA filters yields degraded performance though not statistically significant here. The paradoxical improvement in frame accuracy with an accompanying penalty in WER was consistent in all of our tests with MFSG, including those where we removed the contextual frames to the MLP. The fact that the LDA filters were designed for phone discrimination rather than word recognition, as the original filters were arrived at, may contribute to the discrepancy. This may be further complicated by added non-linear temporal processing inherent in the automatic gain control (AGC) stages in MFSG. It is not uncommon to witness positive gains in frame accuracy leading to negative ones in word recognition and vice versa. An investigation of confusion matrices reveals, interestingly, that MFSG-LDA results in silence being misclassified more often than the original while the correct classification for the phone classes increased. Unfortunately, the direct relation between phone classification and word recognition is difficult to analyze.

5. PLP-LDA

As previously noted, the logarithm of the spectral energies provides an adequate domain with which to successfully apply temporal LDA. Cepstral computation implicitly includes a logarithm followed by a decorrelating linear transformation across frequency bands. The resulting feature components also exhibit class distributions that are approximately normal and thus potentially suitable for LDA. In lieu of frequency-band RASTA filtering we applied the temporal LDA technique to PLP cepstral coefficients. When used in this fashion, the filters are a data-derived replacement of the delta calculations commonly used in ASR systems (figure 3). This manner of LDA use was also recently done by Lieb and Haeb-Umbach using MFCCs in phone recognition tasks [17]. We use this in conjunction with local normalization where the features of each utterance is offset and scaled to zero mean and unit variance. This can be interpreted as a combination of CMS and automatic gain control.

Tables 5 and 6 show WER and frame accuracy results using PLP with no contextual frames available to the MLP probability estimator and 400 hidden units. The direct, delta, and double delta features were replaced with filtering by the first three discriminant components from LDA. We see a similar pattern to our tests with MFSG where we obtain significant improvements in phone classification at the frame level ranging from 5.5% to 17% relative. We also uniformly obtained such frame accuracy improvements in many mismatched training and testing conditions as well. Word recognition improvements unfortunately were not as forthcoming and appeared only in some heavy reverberation tests where the recognition errors remained high.

6. DISCUSSION

Our experiments with a few styles of preprocessing reveal a consistent improvement in phone classification when using temporal LDA. This supports phone recognition results reported in [17]. We also obtained similar results in pilot phone recognition experiments using an unconstrained grammar and two-state minimum duration monophone models. Our MFSG tests with a 100 ms input context to the MLP yielded between 1% and 6% relative phone error reduction when substituting in the LDA filters. Likewise, using LDA in lieu of delta features in our PLP tests with a single frame input to the MLP yielded between 4% and 7% relative improvement. Word recognition tests indicated mixed benefit. Our tests with RASTA-PLP and LDA filters suggest that there is benefit in using temporal filters tuned to the reverberant environment in which the ASR system is trained and tested. This is most apparent in our tests where the MLP probability estimator had access to only a single frame of acoustic features. This advantage is mitigated when we train the MLP with a larger context of adjacent acoustic frames. With 100 ms of acoustic features, the MLP implicitly performs non-linear temporal filtering, the parameters of which are learned through discriminative training. We can effectively consider this a duplication of effort with the discriminative linear filters derived through LDA, though other factors probably exist. One of the motivations for our use of temporal LDA was to insert another level of discriminative training along the temporal dimension in addition to the spectrally related dimension; in a sense, discriminatively spanning the time-frequency plane. With a wide enough sequence of acoustic features, our MLP may already provide an effective coverage. Judicious use of tuned temporal filters may still benefit ASR systems where the allowing a wider acoustic context to the probability estimator becomes prohibitively expensive or

undesirable due to insufficient training data.

In our MFSG and PLP tests, the improved frame accuracy does not produce consistent WER improvements, more frequently the opposite. As the temporal filters were derived through phone discrimination on data, it is not too surprising to witness the improved frame accuracy. Arguably, a good frame accuracy is needed for a good WER, though any further minor improvements through tuning guarantees nothing. It is difficult to analyze how a minor change in phone classification affects the resulting word recognition. Temporal LDA has verified that the modulation frequencies of importance to speech lie in the lower frequencies [6, 10]. It also provides an automatic data-driven means of obtaining relevant filters for phone classification. Directly obtaining filters optimal for word recognition remains elusive and of speculative importance considering the complexity and variety of ASR system implementations and constraints.

7. CONCLUSION

RASTA filtering, CMS, and delta features are forms of temporal processing that have enhanced ASR systems. LDA provides a convenient mechanism for deriving temporal filters that can be used in these processing steps. Further, these filters can be tuned to environments such as reverberation. Our tests show that LDA filters, being designed to discriminate between phones, can consistently improve phone classification. Tests with RASTA-PLP also indicate that this can lead to improvements in word recognition. However, such improvements may be mitigated or made redundant by other means. For example, training an MLP probability estimator with a sufficient acoustic context in reverberation yields comparable results among differently trained LDA filters. Further, word recognition improvements were not forthcoming in tests using alternate preprocessing configurations such as MSFG and PLP-cepstra even while frame accuracy was maintained or improved upon. A common theme with temporal filtering is the enhancement and preservation of the lower modulation frequencies up to about 16 Hz. The further specific parsing of this range, whether by LDA or other means, offers minor tradeoffs among phone classification, word recognition, and performance in reverberant environments.

8. ACKNOWLEDGMENTS

This work was assembled under the kind patronage of Nelson Morgan and ICSI and funded by NSF, grant number (NSF)-IRI-9712579. We wish to express additional appreciation to Hynek Hermansky at OGI for his advice and supervision.

9. REFERENCES

[1] B. S. Atal. Effectiveness of linear prediction characteristics of speech wave for automatic speaker identication and verication. *Journal of the Acoustic Society of America*, (55):1304–12, 1974.

[2] C. Avendano, S. van Vuuren, and H. Hermansky. Data based filter design for RASTA-like channel normalization in ASR. In *ICSLP*, volume 3, pages 2087–90, Philadephia, Pennsylvania, October 1996.

[3] H. Bourlard and N. Morgan. *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Press, 1994.

[4] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. OGI multi-lingual corpus, 1994.

[5] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. Numbers corpus, release 1.0, 1995.

[6] R. Drullman, J. M. Feston, and R. Plomp. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustic Society of America*, 95(2):1053–64, February 1994.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[8] S. Furui. Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transaction on Acoustics Speech and Signal Processing*, (ASSP-34):52, 1986.

[9] S. Greenberg. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop (ETRW) on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32, Pont-a-Mousson, France, 1996. ESCA.

[10] S. Greenberg. Understanding speech understanding: Towards a unified theory of speech perception. In *Proceedings of the ESCA Workshop (ETRW) on The Auditory Basis of Speech Perception*, pages 1–8, Keele, United Kingdom, July 1996. ESCA.

[11] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[12] H. Hermansky. The modulation spectrum in the automatic recognition of speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 140–7, Santa Barbara, CA, USA, December 1997. IEEE Signal Processing Society.

[13] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[14] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel. On the importance of modulation frequencies for speech recognition. In *EUROSPEECH*, volume 3, page 1079, Rhodes, Greece, September 1997.

[15] B. E. D. Kingsbury. *Perceptually Inspired Signal-Processing Strategies for Robust Speech Recognition in Reverberant Environments*. PhD thesis, University of California at Berkeley, 1998.

[16] B. Kinsgbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1-3):117–32, August 1998.

[17] M. Lieb and R. Haeb-Umbach. Lda derived cepstral trajectory filters in adverse environmental conditions. In *ICASSP*, volume 2, pages 1105–8, Istanbul, Turkey, June 2000. IEEE.

[18] N. Morgan and H. Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3):25–42, May 1995.

[19] M. L. Shire and B. Y. Chen. Data-driven rasta filters in reverberation. In *ICASSP*, volume 3, pages 1627–30, Istanbul, Turkey, June 2000. IEEE.

[20] S. van Vuuren and H. Hermansky. Data-driven design of RASTA-like filters. In *EUROSPEECH*, volume 1, pages 1607–1610, Rhodes, Greece, September 1997. ESCA.