

**Auditory Modeling as a Basis for Spectral
Modulation Analysis with Application to
Speaker Recognition**

T.T. Wang
T.F. Quatieri

31 January 2007

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS



Prepared for the Department of Defense under Air Force Contract FA8721-05-C-0002.

Approved for public release; distribution is unlimited.

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense of Defense under Air Force Contract FA8721-05-C-0002.

This report may be reproduced to satisfy needs of U.S. Government agencies.

The ESC Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER


Gary Tutungian
Administrative Contracting Officer
Plans and Programs Directorate
Contracted Support Management

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission has been given to destroy this document when it is no longer needed.

Massachusetts Institute of Technology
Lincoln Laboratory

Auditory Modeling as a Basis for Spectral Modulation Analysis with
Application to Speaker Recognition

T.T. Wang
T.F. Quatieri
Group 62

Technical Report 1119

31 January 2007

Approved for public release; distribution is unlimited.

Lexington

Massachusetts

ABSTRACT

This report explores auditory modeling as a basis for robust automatic speaker verification. Specifically, we have developed feature-extraction front-ends that incorporate (1) time-varying, level-dependent filtering, (2) variations in analysis filterbank size, and (3) nonlinear adaptation. Our methods are motivated both by a desire to better mimic auditory processing relative to traditional front-ends (e.g., the mel-cepstrum) as well as by reported gains in automatic speech recognition robustness exploiting similar principles.

Traditional mel-cepstral features in automatic speaker recognition are derived from ~20 invariant band-pass filter weights, thereby discarding temporal structure from phase. In contrast, cochlear frequency decomposition can be more precisely modeled as the output of ~3500 time-varying, level-dependent filters. Auditory signal processing is therefore more resolved in frequency than mel-cepstral analysis and also derives temporal information. Furthermore, loss of level-dependence has been suggested to reduce human speech reception in adverse acoustic environments. We were thus motivated to employ a recently proposed level-dependent compressed *gammachirp* filterbank in feature extraction as well as vary the number of filters or filter weights to improve frequency resolution. We are also simulating nonlinear adaptation models of inner hair cell function along the basilar membrane that presumably mimic temporal masking effects.

Auditory-based front-ends are being evaluated with the Lincoln Laboratory Gaussian mixture model recognizer on the TIMIT database under clean and noisy (additive Gaussian white noise) conditions. Preliminary results of features derived from our auditory models suggest that they provide complementary information to the mel-cepstrum under clean and noisy conditions, resulting in speaker recognition performance improvements.

TABLE OF CONTENTS

	Abstract	iii
	List of Illustrations	vii
1.	INTRODUCTION	1
2.	AUDITORY MODELING OVERVIEW	3
3.	CEPSTRAL FEATURES WITH IMPROVED FREQUENCY RESOLUTION	7
	3.1 Feature Extraction	7
	3.2 Baseline Recognition Experiments	8
4.	AUDITORY-BASED FEATURE EXTRACTION USING TIME-VARYING NONLINEAR FILTERS	11
	4.1 Auditory-Based Feature Extraction Implementation	11
	4.2 Baseline Recognition Experiments	18
5.	FUTURE EFFORTS	23
	References	25

LIST OF ILLUSTRATIONS

Figure No.		Page
1	Representation of the auditory pathway.	3
2	Auditory model incorporating low- and mid-level auditory processing.	5
3	Equal-energy normalized linear gammatone filters.	8
4	Speaker verification performance with static gammatone filtering with increasing filterbank size. Equal error rate (EER) is shown.	9
5	Composition of gammachirp filter.	12
6	Gammachirp filter with varying levels. Higher levels have lower gain and larger bandwidths.	12
7	Level-dependent filter implementation.	13
8	Example of fast-attack slow-decay estimation of level: (top) input pressure, (bottom) estimated control parameter. Units are normalized.	14
9	Cochleograms derived from (top) mel-, (middle) gammachirp, and (bottom) passive-gammachirp filters for clean speech utterance, “I’ll print out all their records.”	15
10	Same as Figure 9 but with SNR = 30 (left-top), 20 (right-top), 10 (left-bottom), and 0 dB (right-bottom).	16
11	Input envelope (top) and response of Meddis hair cell model simulating adaptation (middle). The log-compression (bottom) with minimum value clipped at ~ -3 normalized amplitude for comparison.	17
12	Equal error rate (EER) performance for the mel-cepstrum with increasing filterbank size in noise (derived from the static gammatone).	18
13	Equal error rate (EER) performance for level-dependent gammachirp filtering with increasing filterbank size in noise.	19

14	(top) Level-dependent gammachirp-filter (ldsc) and mel-cepstrum (mfcc) EER scores for males (m) and females (f) under clean (SNR = 62) and noisy conditions (SNR = 0 – 30); (bottom) performance gains obtained with fusion. Filterbank size = 24 in both methods.	20
15	Performance comparison of Meddis versus DPK nonlinear adaptation, along with score fusion with standard mel-cepstrum.	21

1. INTRODUCTION

Under a variety of adverse acoustic environments, automatic speaker recognition system performance has been shown to degrade (e.g., [Reynolds, 1995]). In contrast, there is evidence that human speech and speaker recognition under similar conditions remain robust [Lippmann, 1997], [SchNielCry, 1998]. Presumably, human auditory processing is able to better extract salient speech information under adverse conditions than standard automatic methods. Indeed, feature extraction methods that mimic peripheral auditory processing have been shown under certain conditions to improve robustness for automatic speech recognition [JanVL, 1995], [TchKoll, 1999].

We hypothesize that more complete auditory modeling may provide a basis for robust speaker recognition. As a first step, we have explored feature extraction methods inspired from peripheral auditory mechanisms that aim to address limitations of traditional front-ends in speaker recognition (e.g., the mel-cepstrum). To the extent that the central auditory system performs spectral modulation analysis, these initial efforts may also provide an improved basis from which to derive such higher-level representations. Preliminary work has accomplished the following:

1. Development of auditory-based feature extraction methods incorporating
 - Variations in analysis filterbank size using auditory-like gammatone filters
 - Time-varying, level-dependent filtering via gammachirp filters
 - Nonlinear adaptation models
2. Baseline experiments in speaker recognition using features derived from our auditory front-end renditions and contrasted against mel-cepstrum features, both in the clear and with additive noise.

In this report, we describe in detail the above accomplishments and outline our future work in further addressing our hypothesis of auditory processing as a basis for robust speaker recognition.

2. AUDITORY MODELING OVERVIEW

Human auditory processing occurs in multiple stages, beginning in the cochlea and auditory nerve of the periphery to the nuclei of the brainstem and midbrain (e.g., cochlear nucleus, inferior colliculus) and subsequently to high-level representations in the auditory cortex [Geisler, 1998][Pickles, 1988] (Figure 1). The cochlea performs mechanical frequency analysis coupled to the summed auditory-nerve outputs from inner hair cells. Physiological evidence has shown that the spectral content of speech is represented in the firing rate and temporal synchrony of auditory nerve firings [Geisler, 1998]. Firing patterns of the auditory nerve are subsequently transmitted through mid-level nuclei to high-level processing centers. Temporal and modulation information has been shown to also be exhibited in these areas [GiraudEtAl, 2000]. We emphasize that in this model the auditory nerve is the sole input to the rest of the auditory pathway; thus, if we aim to exploit higher-level representations (e.g., spectral modulation) similar to those derived in human processing, the peripheral model should be carefully chosen.

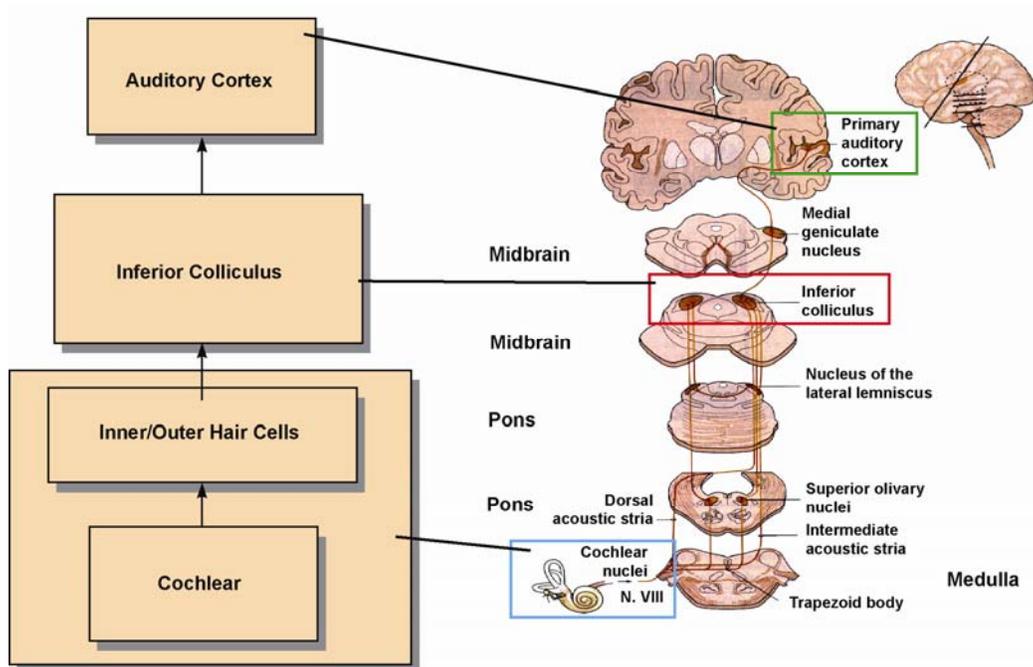


Figure 1. Representation of the auditory pathway [KandelSJ, 2000]

In the traditional mel-cepstrum, spectral magnitude weighting is performed using ~20 band-pass filter weights followed by band-wise energy estimation. In relation to auditory processing, this roughly corresponds to deriving rate information at the auditory nerve after frequency decomposition. A notable discrepancy exists, however, in frequency resolution. Specifically, frequency decomposition in the auditory system is often modeled as the output of ~3500 band-pass filters rather than ~20 filters. It has been suggested that the loss of auditory nerve inputs results in poor speech reception, presumably due to a reduced ability to represent spectral content [SchWool, 1995]. In addition, by performing actual filtering, auditory processing preserves temporal structure associated with signal phase composition while the mel-cepstrum discards it. It has been shown that energy trajectories derived from weighting and filtering are distinct, with those from filtering showing more clearly certain temporal aspects of speech [Quatieri, 2001].

While filter weights of the mel-cepstrum are invariant, auditory filtering is notably nonlinear and time-varying. Physiological and psychophysical evidence has shown that the gain and shape of auditory filters is dependent on input sound level [Moore, 1997]. This nonlinearity presumably results from active mechanisms in the cochlea in response to sound [Geisler, 1998]. It has been suggested that loss of this nonlinearity can account for psychophysical results related to reduced speech reception in adverse acoustic environments [OxenBac, 2003].

We hypothesize that the loss of temporal structure and lack of nonlinearity in the mel-cepstrum limits the robustness of automatic speaker recognition, and were thus motivated to incorporate time-varying, nonlinear filtering in feature extraction. Specifically, our objective at the cochlear level was to implement a recently proposed *gammachirp* filterbank shown to simulate psychophysical and physiological data [IrinoPat, 2006]. In addition, we sought to explore the effects of improving frequency resolution by varying the number of filterbank weights in the context of the traditional cepstral features; this was done using static auditory-like (gammatone) filters. These front-end components are illustrated in Figure 2 with the static gammatone filters, the level-dependent gammachip-filter rendition, and the hair cell model (to be subsequently discussed).

Also shown in Figure 2 are two additional components of our complete auditory model: (1) Synchrony computation that will couple with a complex spectral modulation representation (e.g., the Atlas representation [SchimmAtlas, 2005]) and (2) a modulation filterbank that emulates the mid-level inferior colliculus operation of the auditory pathway [DauKK, 1997]. Ultimately, we will also add to our previously-developed notions on how the high-level auditory regions may process modulation in a two-dimensional time-frequency space [Quatieri, 2002].

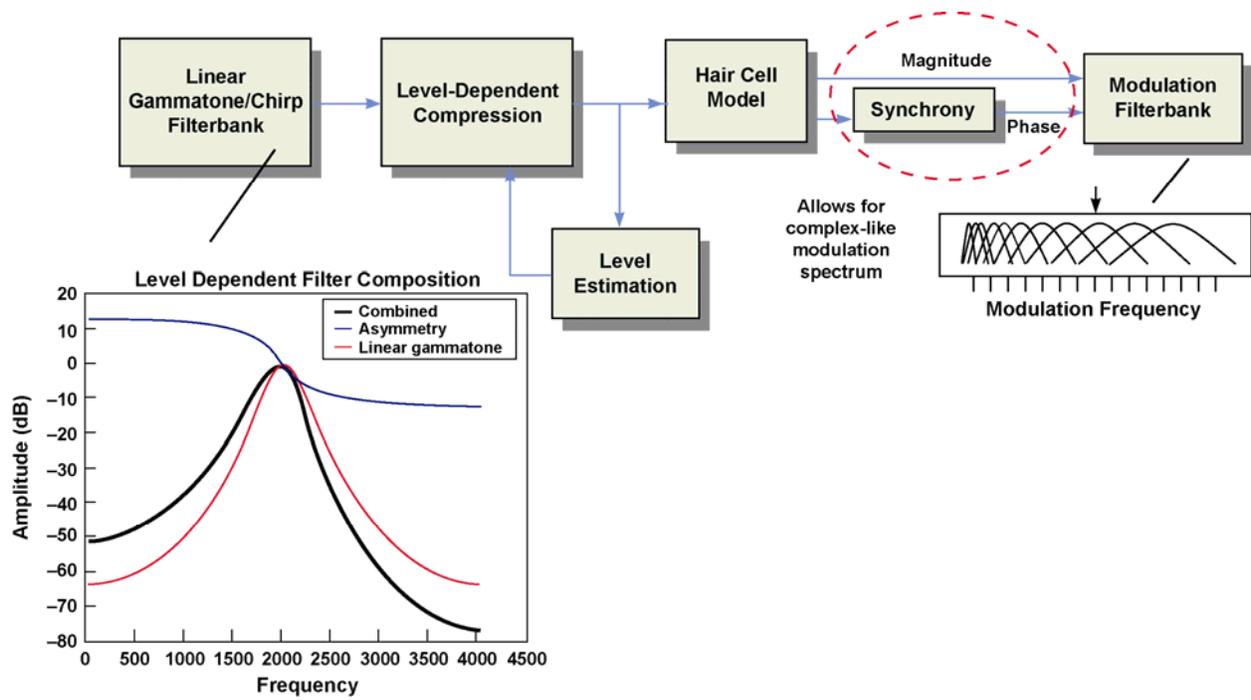


Figure 2. Auditory model incorporating low- and mid-level auditory processing.

3. CEPSTRAL FEATURES WITH IMPROVED FREQUENCY RESOLUTION

Our initial effort was motivated by the possibility that the human auditory system exploits a large number of frequency channels to achieve robustness in speech-related tasks. Specifically, we studied the effects of improving frequency resolution by increasing the number of analysis filter weights used in feature extraction.

3.1 FEATURE EXTRACTION

Speech data were initially pre-emphasized with a frequency-domain weighting function¹. Features were then extracted as in the mel-cepstrum (i.e., spectral magnitude weighting followed by a discrete-cosine transform (DCT) of band-wise log-energies), though with increasing number of filter magnitude weights. We chose this implementation to assess the effects of improved frequency resolution, independent of additional auditory modeling components. Static gammatone filterbank weights were initially used and covered the same frequency region of the standard (mel-cepstrum) triangular filters [Slaney, Toolkit] (Figure 3). We then generalized this static gammatone set from ~24 filters to an arbitrary number by sampling a nonlinear continuous function representing the logarithmic relation between basilar membrane location and center frequencies and their bandwidths. Because the feature dimensionality grows with increasing filter number, we explored two different approaches:

1. Allow the dimensionality to *grow* with increasing filterbank size.
2. *Trim* the dimensionality by keeping only the first 19 DCT coefficients (excluding the 0th coefficient).

¹ The pre-emphasis function is given by $P(f) = 1 + f^2 / 25000$ with f being continuous-time frequency.

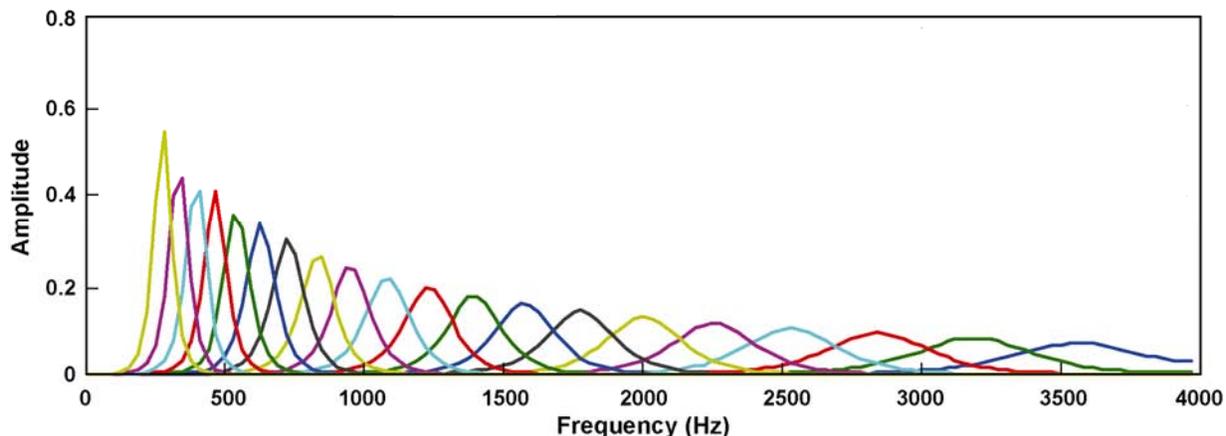


Figure 3. Equal-energy normalized linear gammatone filters.

3.2 BASELINE RECOGNITION EXPERIMENTS

Preliminary results (Figure 4) in varying gammatone filterbank size do not show a significant correlation between number of filters and speaker recognition performance. We used in this initial experiment the TIMIT corpus with 256 target speakers [FisherDG, 1986]. TIMIT was recorded in the clear at a 16k sampling rate, is phonetically balanced, and has a good representation of speaker types and dialects. We chose TIMIT as our initial testing corpus in order to isolate the importance of the candidate features themselves without influence from uncontrolled background or channel distortion.

The classifier used is the Lincoln GMM-UBM system, and no channel or score compensation is applied [ReyQuatDunn, 2002]. The background model was trained with 168 TIMIT speakers, equally distributed with males and females, with 10 TIMIT utterances for each speaker. There were 137 female and 326 male target speakers, with eight training utterances and two test utterances for each. In these experiments, we used the above trimmed and untrimmed versions of our DCT coefficients, giving 19 coefficients (trimmed) and $N = \text{filterbank size} - 1$ (untrimmed), respectively. The general equal error rate (ERR) performance in these two cases (Figure 4), both for males and females, is close to that of the standard 19th-order mel-cepstrum features (not shown).

Though this result may reflect the sufficiency of frequency resolution of the mel-cepstrum, we do not believe this to be true. With trimming, the lack of performance gains may stem from our method of feature reduction. Without trimming, the number of filters corresponds to the total number of features derived in cepstral analysis. In the first case, we sought to avoid the “curse of dimensionality” in pattern recognition by using only the first 19 coefficients (excluding the zeroth value) as in the DCT. It is unclear whether this “trimming” of coefficients effectively exploits the large number of channels available. With

no trimming, although more speech information is available, we may have encountered the curse of dimensionality, in particular, for a GMM-based classifier. Other classifiers, such as support vector machines, may thwart this dimensionality tradeoff [CampSRS, 2006].

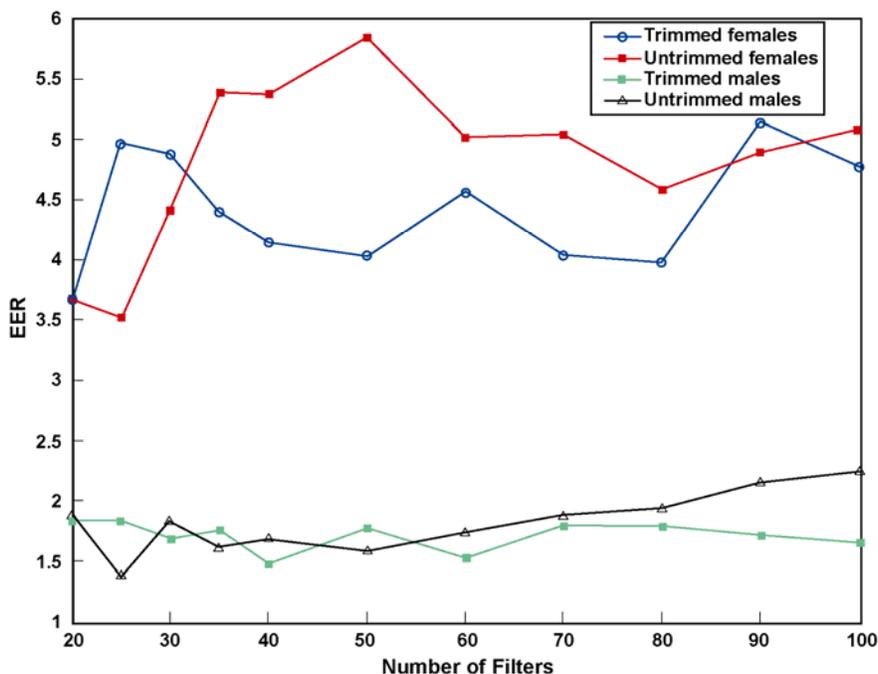


Figure 4. Speaker verification performance with static gammatone filtering with increasing filterbank size. Equal error rate (EER) is shown.

It is interesting to observe in Figure 4 the performance gap between males and females. This gap has also been found in previous studies with the use of mel-cepstrum features in speaker recognition. A speculation was that by increasing the size of the front-end filterbank, we might reduce the effect of aliasing in the DCT, occurring for high pitch with sparse spectral sampling, and thus with trimming improve the accuracy of low-DCT coefficients. It appears, however, that under-sampled formant information cannot be recovered by a single-frame representation, and that alternate methods such as exploiting temporal signal modulation may be the key to reducing this gender gap.

4. AUDITORY-BASED FEATURE EXTRACTION USING TIME-VARYING NONLINEAR FILTERS

In the previous section, we studied the effects of improving frequency resolution using the static component (i.e., the gammatone) of the level-dependent gammachirp for spectral magnitude weighting and reported on baseline speaker recognition experiments. In this section, we consider a more complete peripheral auditory model by including the chirp component of the gammachirp filterbank, level-dependent compression, and two different hair cell models, one by Dau et al. [DauPK, 1996] and the other by Meddis [Meddis, 1986]. In addition, we perform a variety of baseline speaker recognition experiments using features derived from these various auditory models, both in the clear and in noise. Based on our recognition results from Section 3.2, when investigating filterbank size, we will invoke the trimmed case only.

4.1 AUDITORY-BASED FEATURE EXTRACTION IMPLEMENTATION

Model components were assessed individually and in combination for feature extraction. Pre-emphasis was initially performed on speech data using a first-difference operation².

Level-dependent filtering: Distinct from the previous section, features from the auditory model are derived from actual filtering rather than spectral magnitude weighting (see Section 2). The gammachirp filter that we have implemented is composed of three cascaded infinite-impulse-response (IIR) filters: a linear gammatone filter (from our previous section), a passive low-pass asymmetry, and a level-dependent asymmetry (Figure 5). The level-dependent asymmetry with the passive asymmetry modifies the shape and gain of the filter based on input level (Figure 6); with increasing input levels, gains are reduced while bandwidths are increased.

² The pre-emphasis function in discrete time is given by $y[n] = x[n] - 0.97x[n-1]$.

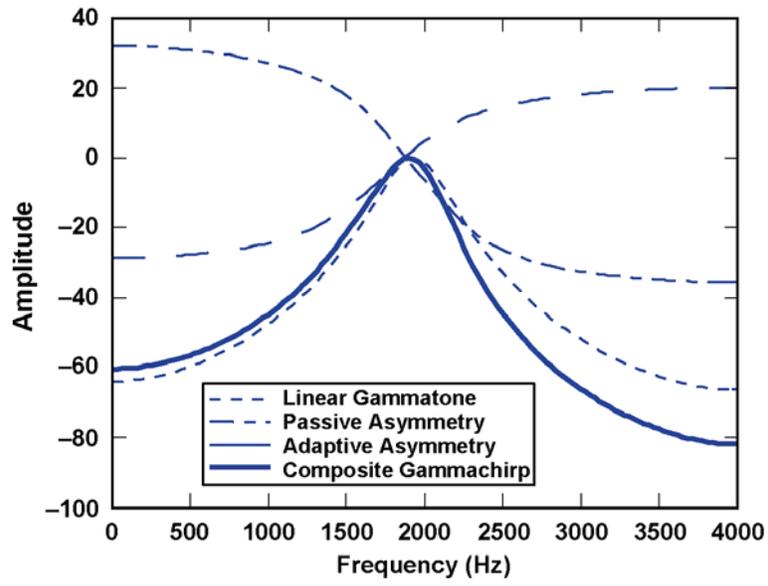


Figure 5. Composition of gammachirp filter.

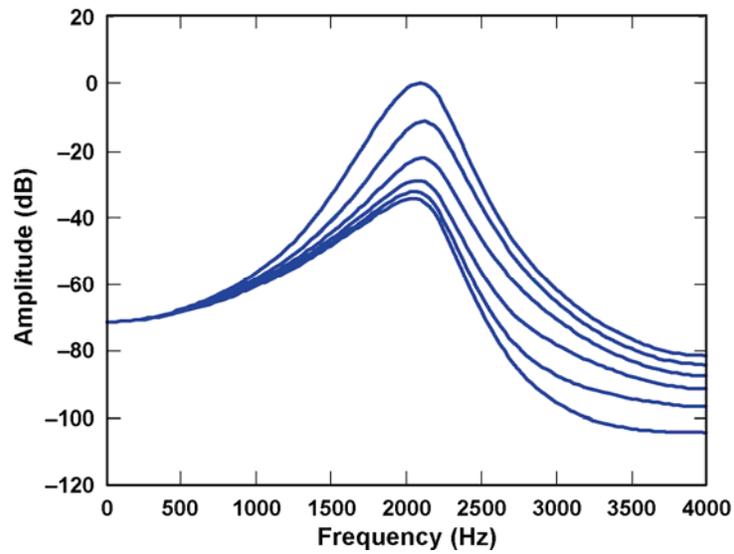


Figure 6. Gammachirp filter with varying levels. Higher levels have lower gain and larger bandwidths.

In filtering, the signal is filtered through parallel paths to determine the filtered result while estimating the input level (Figure 7). Level-estimation is done using a *fast-attack slow-decay* function that follows an input level to a local maximum and then decays according to a half-life parameter (Figure 8) [IrianoPat, 2006].

For each feature case, the envelope of the filtered speech signal from each auditory channel (derived from the Hilbert transform³) is low-pass smoothed and down-sampled, and is followed by the DCT analysis. The low-pass filter cutoff used in envelope smoothing is set consistent with downsampling to a 10-ms frame interval. Although smoothing reduces temporal resolution, it is essential to avoid aliasing in a feature representation.

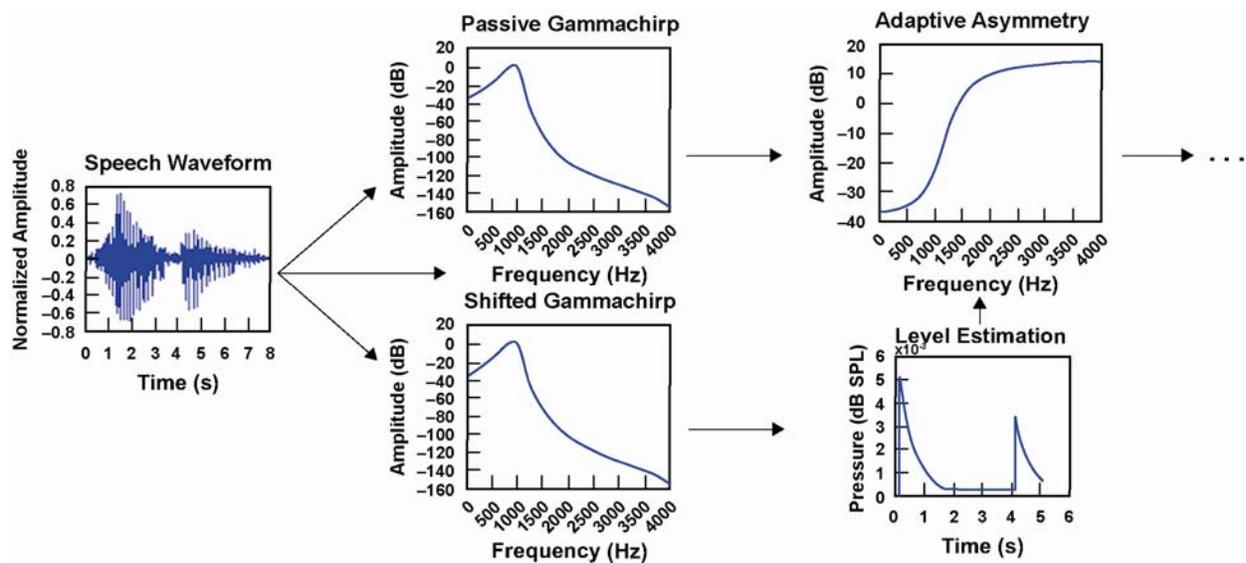


Figure 7. Level-dependent filter implementation.

³ Other possibilities exist for envelope calculation. For example, rectification and low-pass filtering may be closer to actual auditory processing.

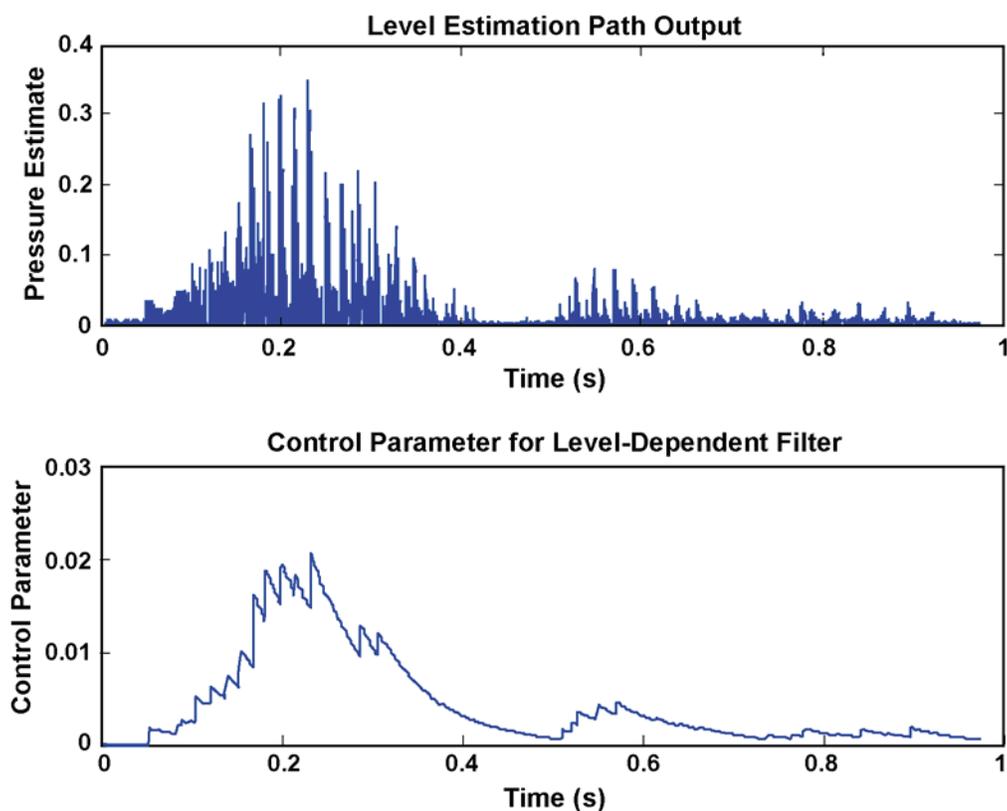


Figure 8. Example of fast-attack slow-decay estimation of level: (top) input pressure, (bottom) estimated control parameter. Units are normalized.

Cochleograms⁴ derived using the gammachirp filter have been suggested to accentuate salient features of speech [IrinPat, 2006]. Our preliminary work has compared the outcome of the gammachirp filters with traditional mel-cepstral filters. Figures 9 and 10 show cochleograms derived using 1) energy-equalized mel-cepstral filters, i.e., standard triangularly-shaped static filters, 2) gammachirp filters (with level-dependence), and 3) energy-equalized passive gammachirp filters (without level-dependence, i.e., the static gammatone case) for a speech utterance in the presence of additive Gaussian white noise for several signal-to-noise ratios (SNR) and under a clean condition. Responses are similar under clean and near-clean conditions, while the gammachirp notably exhibits less relative distortion in higher frequency channels as SNR decreases. Although these results appear promising for robust feature extraction, a

⁴A correlogram is a time-frequency distribution derived from output envelopes of an auditory-like filterbank.

significant caveat is that differences in distortion were not present for all of the speech utterances examined. Further work is needed in examining the mechanisms leading to these differences to assess how they may be exploited for robust feature extraction.

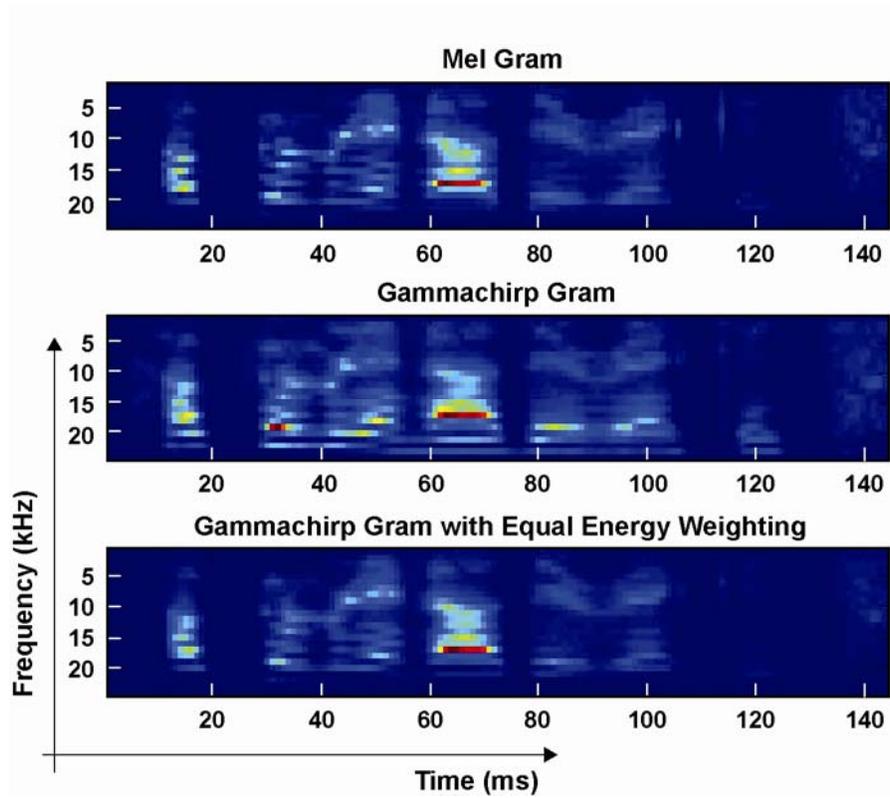


Figure 9. Cochleograms derived from (top) mel-, (middle) gammachirp, and (bottom) passive-gammachirp filters for clean speech utterance, "I'll print out all their records."

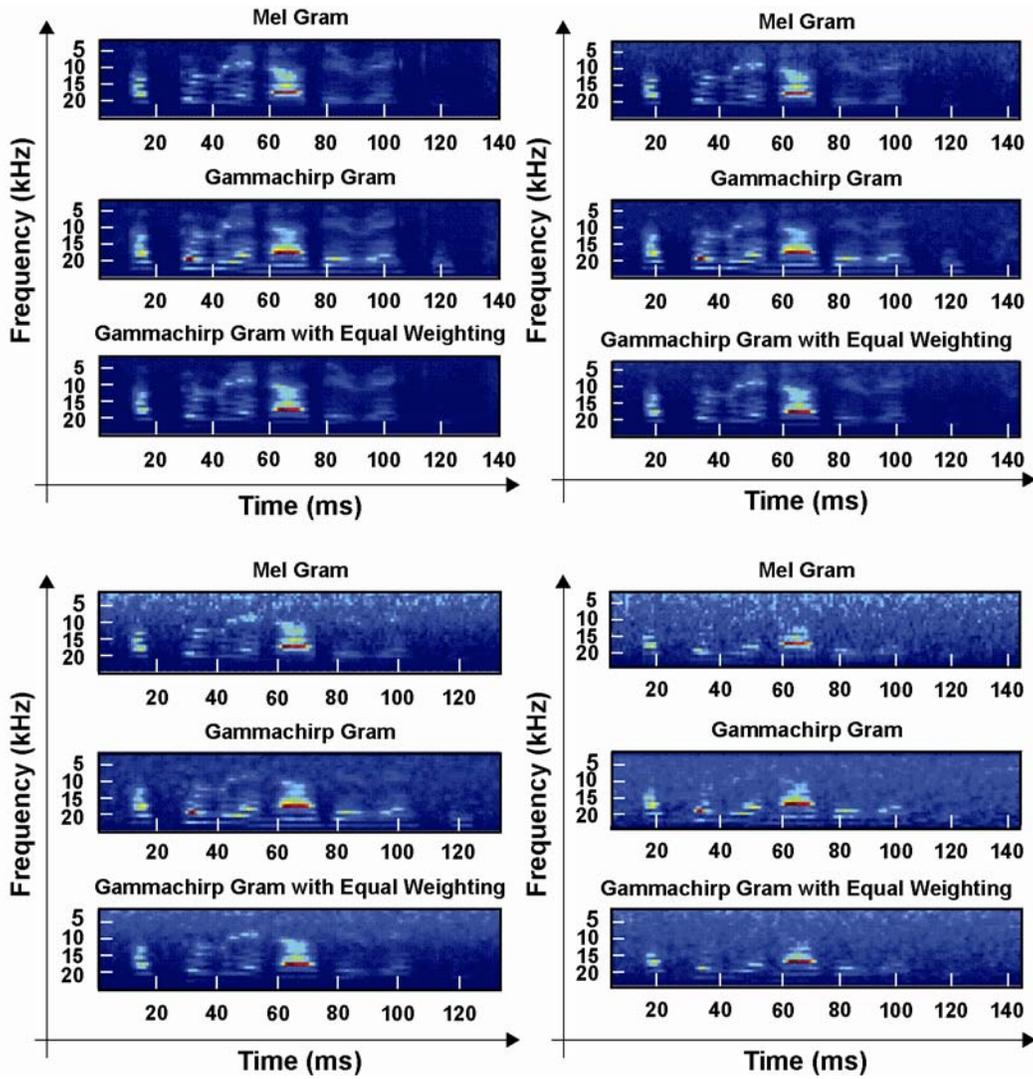


Figure 10. Same as Figure 9 but with SNR = 30 (left-top), 20 (right-top), 10 (left-bottom), and 0 dB (right-bottom).

Nonlinear adaptation: Activity at the auditory-nerve level, allowing spectral and temporal representations of speech, results from the synaptic release processes of the inner hair cell [Geisler, 1998]. The dynamics of these processes results in unique temporal characteristics of auditory nerve activity collectively referred to as *adaptation*. As observed for sustained stimuli (e.g., a pure tone), post-stimulus time histograms of auditory-nerve activity exhibit sharp onsets followed by compression; furthermore, at the stimulus offset, an overshoot of reduced activity is present (Figure 11). In the context of feature

extraction from the envelope of an analysis filterbank, a notable distinction between nonlinear adaptation and the log function used in cepstral analysis is the enhancement of transient behavior (Figure 11) [QuatMS, 2003]. Furthermore, it is likely that adaptation is responsible for the psychophysical phenomenon of masking. Reported gains in automatic speech recognition (ASR) noise robustness when this nonlinearity is incorporated in feature extraction are presumably the result of these distinctions [TchKoll, 1999].

In combination with the gammachirp filterbank, we have evaluated two models of nonlinear adaptation for use in feature extraction for speaker recognition: The first is the inner hair cell model by Dau, Puschel, and Kohlrausch (DPK) [DauPK, 1996] and the second by Meddis [Meddis, 1986]. The DPK model is derived from psychophysical results in forward masking. The Meddis model is based on purported mechanisms of synaptic transmission and is fit to physiological data. Acting on the envelope from the gammachirp filter output, both models are used in place of the log function in cepstral analysis; all other aspects of feature extraction are identical to those discussed in the following section.

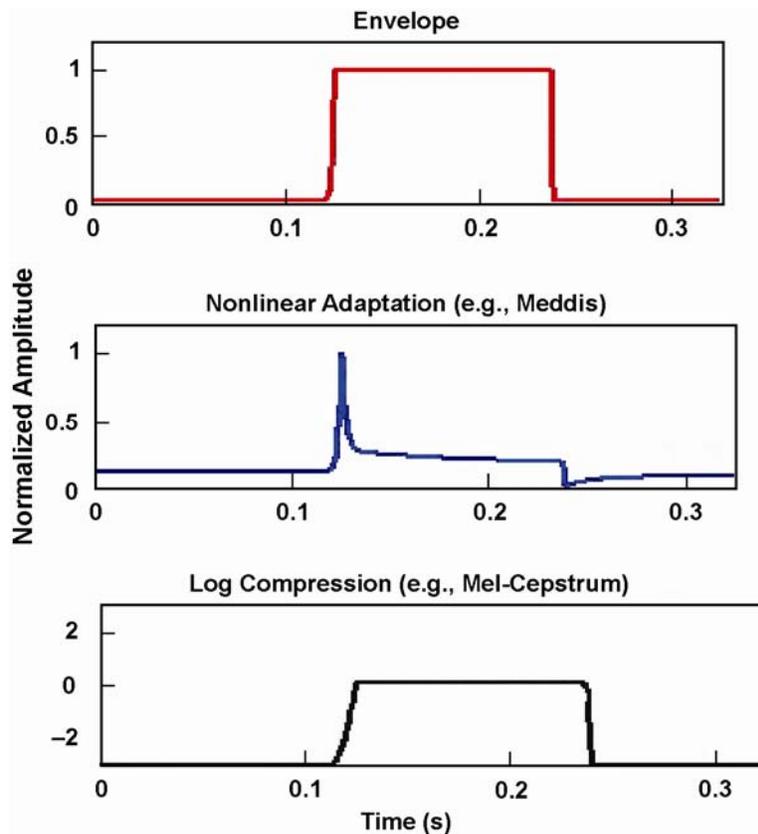


Figure 11. Input envelope (top) and response of Meddis hair cell model simulating adaptation (middle). The log-compression (bottom) with minimum value clipped at ~ -3 normalized amplitude for comparison.

4.2 BASELINE RECOGNITION EXPERIMENTS

In our speaker recognition experiments, we again used the TIMIT corpus. Experimental setup is the same as that given in Section 3.2 for the clean condition. For noisy conditions, Gaussian white noise was added to training and testing data. Feature extraction involved (1) increasing filterbank size in noise with the level-dependent gammachirp alone, and (2) inclusion of the DPK and Meddis adaptation models with the level-dependent gammachirp. Resulting scores are compared to the mel-cepstrum.

Preliminary equal-error-rate (EER) results in varying filterbank size in noise are shown in Figures 12 and 13. Mel-cepstral features are derived using the traditional frequency weighting pre-emphasis $P(f) = 1 + f^2 / 25000$ while the gammachirp employs the first-difference ($y[n] = x[n] - 0.97x[n-1]$). Both feature sets are derived from the static gammatone weights (see Section 3). We used the trimmed versions of our DCT coefficients, giving 19 coefficients, which is smaller than the full DCT length. The technique of trimming was introduced in the previous section. For this scenario, our preliminary results do not show a significant correlation between number of filters and performance for both mel-cepstral weighting and level-dependent gammachirp-filtering front-ends. Furthermore, results show similar performance between the mel-filter and gammachirp cases, although the mel-cepstrum overall performs slightly better.

As we argued for the static gammatone results, we believe the results of the gammachirp reflect insufficiency in our dimensionality reduction method of “trimming” rather than the sufficiency of the mel-cepstrum’s frequency resolution.

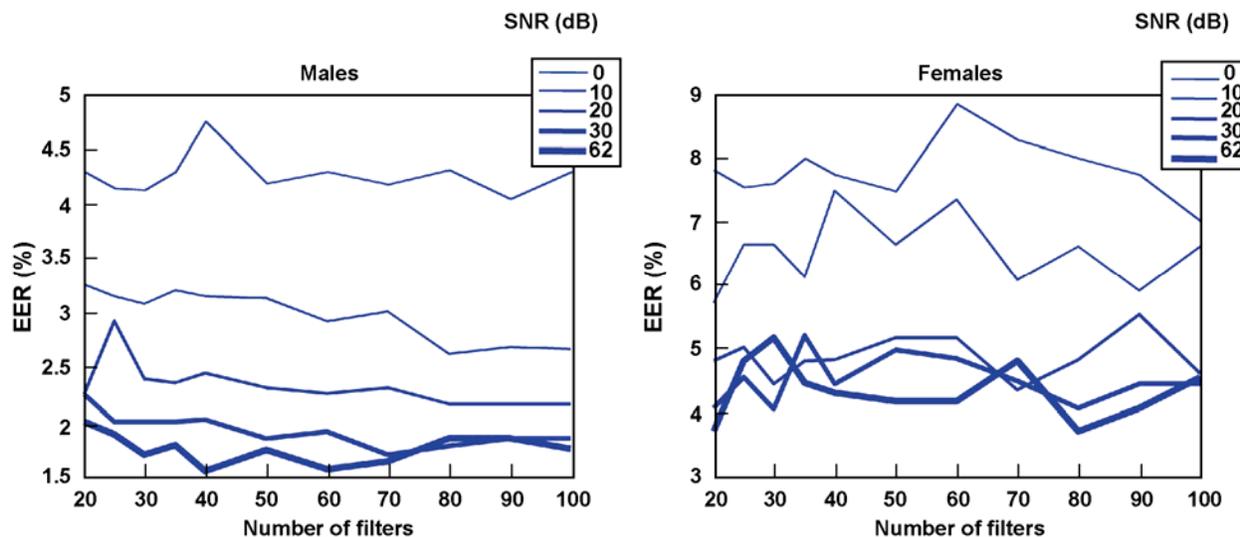


Figure 12. Equal error rate (EER) performance for the mel-cepstrum with increasing filterbank size in noise (derived from the static gammatone).

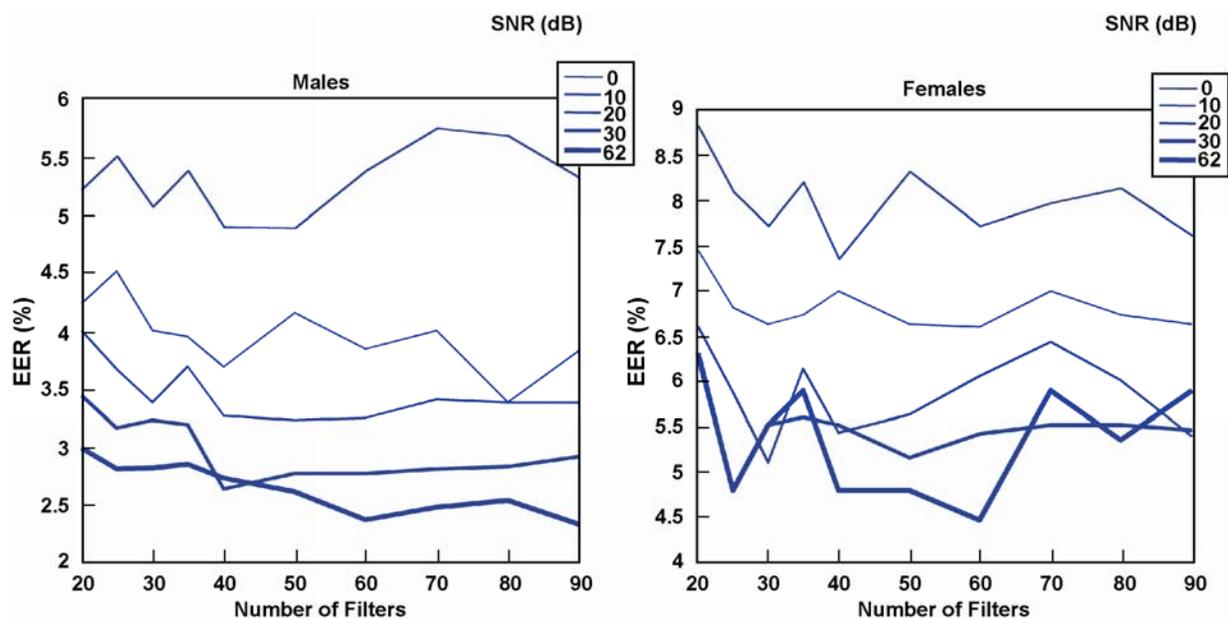


Figure 13. Equal error rate (EER) performance for level-dependent gammachirp filtering with increasing filterbank size in noise.

For a stricter comparison, Figure 14 shows results from level-dependent gammachirp filtering with the mel-cepstrum derived from traditional triangular critical-band weights. Pre-emphasis is the same in both cases, employing the first-difference. Filterbank size was set to 24 with trimming to 19 DCT coefficients in both cases, as is typically done in the mel-cepstrum. We obtain reductions in EER ranging from 0.5~2% with fusion of the scores (equal weighting) with greater gains at lower SNR. The comparable performance of the gammachirp with the mel-cepstrum and the gains obtained with fusion suggest that the gammachirp filtering may provide complementary information to the mel-cepstrum.

Finally, our initial comparison of nonlinear adaptation models shows that the Meddis model consistently outperforms the DPK in speaker verification, though both are less robust than the mel-cepstrum. The comparative EER results are shown in Figure 15 for males, along with the result of score-fusion of Meddis- and mel-cepstrum-based features (showing negligible fusion gain). Similar relative gains hold for females. That the Meddis model is higher performing than the DPK is perhaps due to inherent weaknesses of the DPK as a purely phenomenological model. Specifically, it has been shown that its response to transient behavior and its choice of parameters are physiologically implausible⁵. The performance gap between the mel-cepstrum and the Meddis model may be attributed to limitations of our

⁵ Thanks to Nicolas Malyska for a personal communication.

current feature extraction methods in mimicking auditory processing, assuming of course that auditory processing is our “gold standard.” Most notably, the Meddis model is derived from physiological responses to *acoustic* stimulation. Nonlinearities in auditory filtering are therefore already incorporated via the model’s choice of parameters. By acting on the envelope of the gammachirp, filtering nonlinearities are likely twice represented, thereby distorting the auditory representation of the signal.

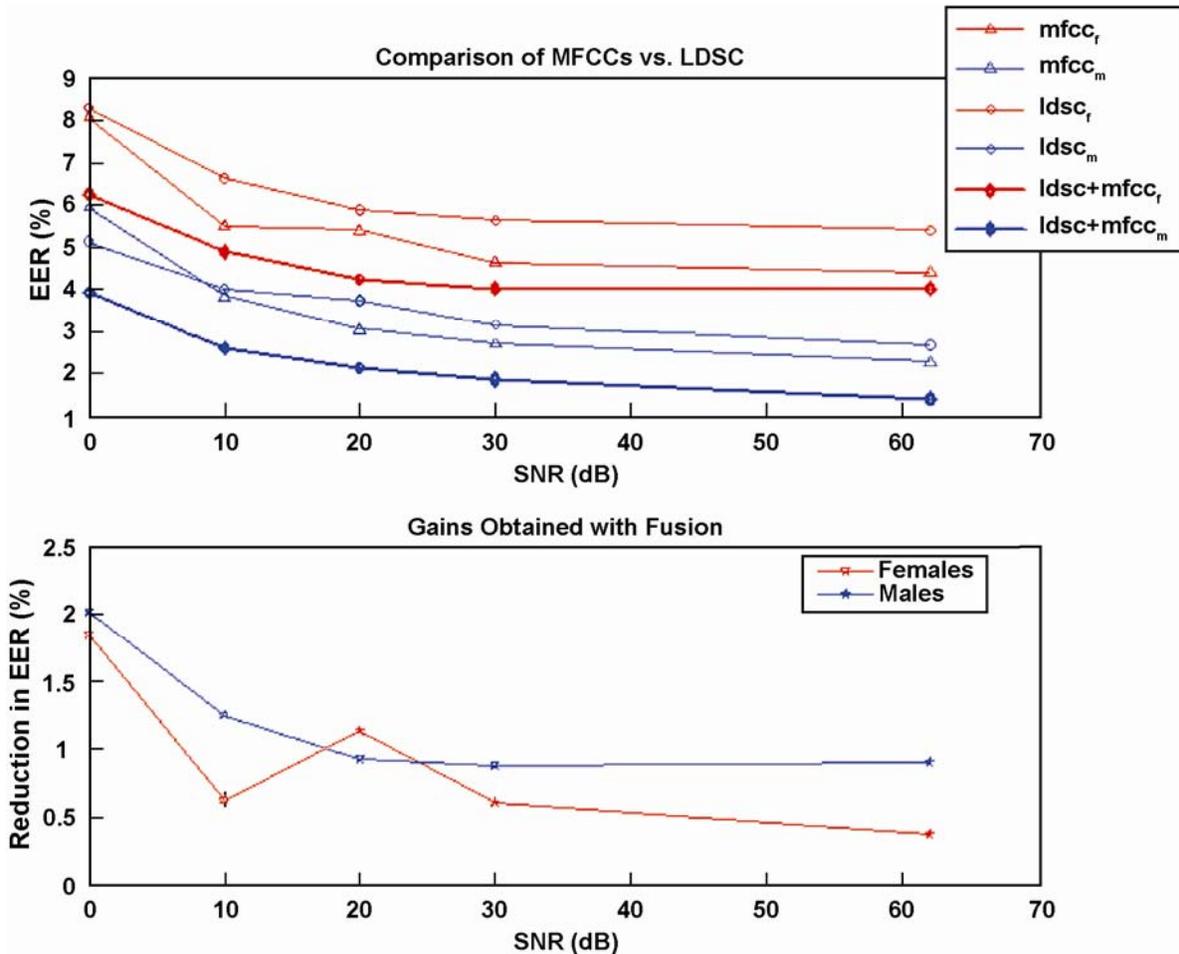


Figure 14. (top) Level-dependent gammachirp-filter (*ldsc*) and mel-cepstrum (*mfcc*) EER scores for males (*m*) and females (*f*) under clean ($SNR = 62$) and noisy conditions ($SNR = 0 - 30$); (bottom) performance gains obtained with fusion. Filterbank size = 24 in both methods.

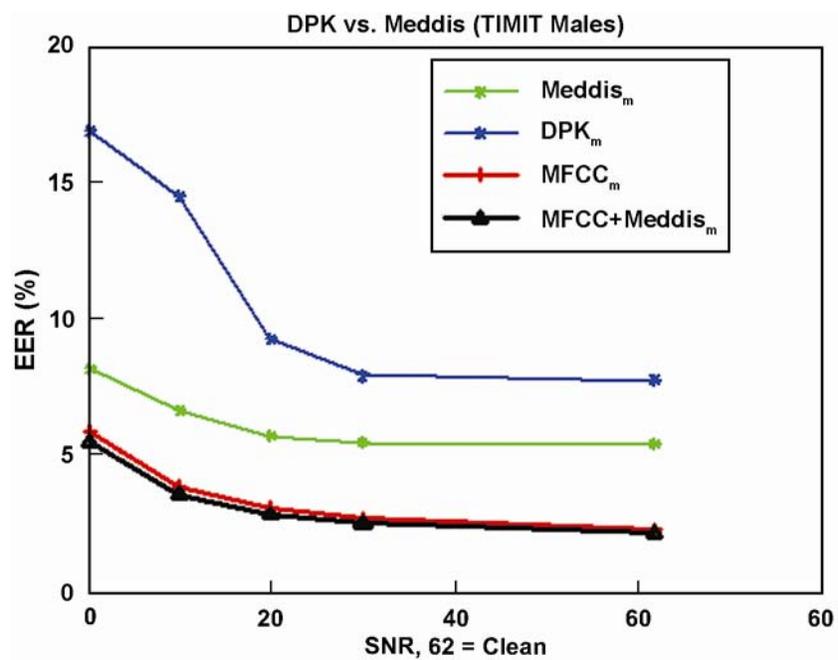


Figure 15. Performance comparison of Meddis versus DPK nonlinear adaptation, along with score fusion with standard mel-cepstrum.

5. FUTURE EFFORTS

In summary, preliminary results suggest that incorporation of level-dependent filtering in feature extraction may provide complementary information to the mel-cepstrum in noise. Nevertheless, we have only touched upon the use of auditory modeling for speaker recognition. Our work has raised a multitude of questions to be answered in future work.

We will explore a number of approaches to better understand how auditory processing results in robust representations of speech in noise. For example, we have observed that gammachirp filtering is able to reduce the effects of noise for certain speech segments but not others. One possible explanation for this limitation is the “open-loop” nature of the model, unlike the actual “closed-loop” periphery consisting of descending efferent feedback from central pathways. Both physiological and psychophysical evidence suggest that efferent stimulation plays a role in noise suppression. Stimulation reduces auditory filter gains and enhances auditory nerve responses to transients in noise [KawaseEtAl, 1993]. In addition, speech-in-noise reception benefits from activation of efference in humans [GiraudEtAl, 1997]. Incorporation of a control component similar to efference with gammachirp filtering may therefore allow more robustness in our auditory representations.

Our current frontend also utilizes the envelope from filtering to derive the speech spectrum while discarding potential contributions from temporal synchrony (i.e., auditory nerve phase-locking). Synchrony-derived representations, however, have been shown to be more robust in noise for automatic speech recognition than rate-based methods (e.g., [Ghitza, 1987]). Our future work will explore the explicit incorporation of neural synchrony as a means to improve robustness.

Limitations of our current front-end will also be addressed in the general scope of speaker recognition. Specifically, incorporation of models that exclusively mimic synaptic transmission (e.g., [SumnerLPMM, 2002]) with gammachirp filtering would eliminate the nonlinear redundancy previously noted and better mimic auditory processing. To better exploit larger filterbanks, two possible directions are to utilize traditional dimensionality reduction methods (e.g., principle components analysis) or to employ machine learning paradigms shown to be robust to large feature size (e.g., support vector machines [CampSRS, 2006]). Alternatively, outputs of peripheral filtering could be used in conjunction with models of higher processing centers in the auditory pathway in the reduction of dimensionality, as has been suggested by Yang, et al. [YanWanSha, 1992].

Our future work will also involve integration of our auditory-based approach with a complex modulation envelope approach (such as that by Atlas [SchimmAtlas, 2005]) by introducing neural synchrony, Ghiza’s modulation bandwidth paradox [Ghitza, 2001], and Dau et al.’s modulation filterbank interpretation of the inferior colliculus [DauKK, 1997]. Ultimately, we will also introduce two-dimensional processing speculated to occur in high-level auditory mechanisms [Quatieri, 2002]. In all future work, although some testing will be done in the clear, our emphasis will be on furthering our understanding of auditory processing of modulation in noise environments and testing our feature extraction in these environments.

REFERENCES

- [CampSRS, 2006] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff. “SVM-based speaker verification using a GMM supervector and NAP variability compensation,” *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2006.
- [DauKK, 1997] T. Dau, B. Kollmeier, and A. Kohlrausch, “Modeling auditory processing of amplitude modulation.1. Detection and masking with narrow-band carriers,” *J. Acoust. Soc. Am.*, vol. 102, pp. 2892-2905, 1997.
- [DauPK, 1996] T. Dau, D. Puschel, and A. Kohlrausch, “A quantitative model of the ‘effective’ signal processing in the auditory system: I. Model structure,” *J. Acoust. Soc. Am.*, vol. 99, 3615-3622, 1996.
- [FisherDG, 1986] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, “The DARPA speech recognition research database: specifications and status,” *Proc. of the DARPA Speech Recognition Workshop*, 1986.
- [Geisler, 1998] C.D. Geisler, *From sound to synapse*, Oxford University Press, Oxford, 1998.
- [Ghitza, 1987] O. Ghitza, “Robustness against noise: The role of timing-synchrony measurement,” *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, April 1987.
- [Ghitza, 2001] O. Ghitza, “On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception,” *J. Acoust. Soc. Am.*, vol. 110 (3), Pt. 1, September 2001.
- [GiraudEtAl, 1997] A-L. Giraud, S. Garnier, C. Micheyl, G. Lina, A. Chays, S. Chery-Croze, “Auditory efferents involved in speech-in-noise intelligibility,” *Neuroreport*, 8: 1779 – 1783, 1997.
- [GiraudEtAl, 2000] A-L. Giraud, C. Lorenzi, J. Ashburner, J. Wable, I. Johnsrude, R. Frackowiak, and A. Kleinschmidt, “Representation of the Temporal Envelope of Sounds in the Human Brain,” *Journal of Neurophysiology* 84: 1588 – 1598, 2000.
- [IrinoPat, 2006] T. Irino and R.D. Patterson, “A dynamic, compressive gammachirp auditory filterbank,” *Trans. of IEEE Acoustic,s Speech, and Language Processing*, Oct 2006.
- [JanVL, 1995] C. Jankowski, H. Vo, and R. Lippmann, “A comparison of signal processing frontends for automatic word recognition,” *Trans. Of IEEE Acoustics Speech and Language Processing*, vol. 3, no. 4, 1995.
- [KandelSJ, 2000] E.R. Kandel, J.H. Schwartz, and T.M. Jessel (editors), *Principles of Neural Science*, Fourth Edition, McGraw Hill, New York, NY, 2000.

- [KawaseEtAl, 1993] T. Kawasi, M.C. Liberman, B. Delgutte, "Antimasking effects of the olivocochlear reflex II: Enhancement of auditory-nerve response to masked tones," *Journal of Neurophysiology*, vol. 70, no. 6, 1996.
- [Lippmann, 1997] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, 22, 1-15, 1997.
- [Meddis, 1986] Meddis, R. "Simulations of mechanical to neural transduction in the auditory receptor," *J.Acoust.Soc.Amer.*, vol. 79, no. 3, 702-711, 1986.
- [Moore, 1997] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, San Diego, CA, 1997.
- [OxenBac, 2003] A.J. Oxenham and S.P. Bacon, "Cochlear compression: Perceptual measures and implications for normal and impaired hearing," *Ear and Hearing*, 24, 352-366, 2003.
- [Pickles, 1988] J.O. Pickles, *An Introduction to the Physiology of Hearing*. London, UK: Academic, 1988.
- [Quatieri, 2001] T.F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Upper Saddle River, NJ: Prentice Hall, 2001.
- [QuatMS, 2003] T. F. Quatieri, N. Malyska, and D. Sturim, "Auditory signal processing as a basis for speaker recognition," *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003.
- [Quatieri, 2002] T.F. Quatieri, "Two-dimensional processing of speech with application to pitch estimation," *Proc. of Int. Conf. Speech and Language Processing*, 2002.
- [Reynolds, 1995] D.A. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46-48, March 1995.
- [ReyQuatDunn, 2002] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, Special Issue: NIST 1999 Speaker Recognition Workshop, Academic Press, vol. 10, no. 1-3, pp. 19-41, January/April/July 2000.
- [SchNielCry, 1998] A. Schmidt-Nielsen and T.H. Crystal. "Human vs. machine speaker identification with telephone speech," *Proc. of Int. Conf. Speech and Language Processing*, December 1998.
- [SchWoel, 1995] H.F. Schuknecht and R.C. Woellner, "An experimental and clinical study of deafness from lesions of the cochlear nerve," *J. Laryngol. Otol.* 69: 75-97, 1995.

[SchimmAtlas, 2005] S. Schimmel and L. Atlas, “Coherent envelope detection for modulation filtering of speech,” *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2005.

[Slaney, Toolkit] M.Slaney, <http://www.slaney.org/malcolm/pubs.html>.

[SumnerLPMM, 2002] C. J. Sumner, E. A. Lopez-Poveda, L. P. O’Mard, and R. Meddis, “A revised model of the inner-hair cell and auditory-nerve complex,” *J. Acoust. Soc. Am.*, vol. 111, pp. 178–88, May 2002.

[TchKoll, 1999] J. Tchorz and B. Kollmeier, “Auditory-Based Feature Extraction for Robust Speech Recognition in Noisy Environment,” *J. Acoust. Soc. Am.* pp 2040 – 2050, 1999.

[YanWanSha, 1992] X.W. Yang, K.S. Wang, S.A. Shamma, “Auditory Representations of Acoustic Signals,” *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824 – 839, March 1992.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE 31 January 2007		2. REPORT TYPE Technical Report		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Auditory Modeling as a Basis for Spectral Modulation Analysis with Application to Speaker Recognition				5a. CONTRACT NUMBER FA8721-05-C-0002	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Tianyu Tom Wang and Thomas F. Quatieri				5d. PROJECT NUMBER 1306	
				5e. TASK NUMBER 0	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) MIT Lincoln Laboratory 244 Wood Street Lexington, MA 02420-9108				8. PERFORMING ORGANIZATION REPORT NUMBER TR-1119	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Department of Defense				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) ESC-TR-2006-076	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT This report explores auditory modeling as a basis for robust automatic speaker verification. Specifically, we have developed feature-extraction front-ends that incorporate (1) time-varying, level-dependent filtering, (2) variations in analysis filter-bank size, and (3) nonlinear adaptation. Our methods are motivated both by a desire to better mimic auditory processing relative to traditional front-ends (e.g., the mel-cepstrum) as well as by reported gains in automatic speech recognition robustness exploiting similar principles. Traditional mel-cepstral features in automatic speaker recognition are derived from ~20 invariant band-pass filter weights, thereby discarding temporal structure from phase. In contrast, cochlear frequency decomposition can be more precisely modeled as the output of ~3500 time-varying, level-dependent filters. Auditory signal processing is therefore more resolved in frequency than mel-cepstral analysis and also derives temporal information. Furthermore, loss of level-dependence has been suggested to reduce human speech reception in adverse acoustic environments. We were thus motivated to employ a recently proposed level-dependent compressed <i>gammachirp</i> filter bank in feature extraction as well as vary the number of filters or filter weights to improve frequency resolution. We are also simulating nonlinear adaptation models of inner hair cell function along the basilar membrane that presumably mimic temporal masking effects. Auditory-based front-ends are being evaluated with the Lincoln Laboratory Gaussian mixture model recognizer on the TIMIT database under clean and noisy (additive Gaussian white noise) conditions. Preliminary results of features derived from our auditory models suggest that they provide complementary information to the mel-cepstrum under clean and noisy conditions, resulting in speaker recognition performance improvements.					
15. SUBJECT TERMS auditory modeling nonlinear adaptation robust automatic speaker verification level-dependent filtering gammachirp filter-bank Gaussian mixture model					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as report	18. NUMBER OF PAGES 36	19a. NAME OF RESPONSIBLE PERSON
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (include area code)

