Using Fast and Slow Modulations to Model Human Hearing of Fast and Slow Speech

Shuo-Yiin Chang[1,2], Nelson Morgan[1,2], Anirudh Raju[3,4], Abeer Alwan[3], and Jody Kreiman[5]

TR-15-003

August 2015

Abstract

A collaboration between the Speech Processing and Auditory Perception laboratory at UCLA and the Speech Group at ICSI focused on the refinement of the simple models used in ASR with representations that have been filtered in the modulation domain to better match human perception. To quantitatively measure the effects of this modification, UCLA collected CVC stimuli uttered quickly and more slowly, and conducted perceptual tests for clean and noisy versions of the stimuli. The ICSI team then conducted tests to determine if inclusion of Gabor-filtered spectrograms with lower or higher temporal modulations could be used to correlate better with human perception. Here we report on results that confirmed an improvement in this correlation, particularly for noisy and rapid speech, while also improving the accuracy. Overall accuracies in noise for all systems tested, though, were quite poor, suggesting that further auditory modeling might be necessary to improve the modeling of human performance on this task.

1 International Computer Science Institute, Berkeley, CA, USA
2 EECS Department, UC Berkeley, Berkeley, CA, USA
3 Electrical Engineering Department, UC Los Angeles, Los Angeles, CA, USA
4 Amazon.com, Inc., Seattle, WA, USA
5 Department of Bioengineering, UC Los Angeles, Los Angeles, CA, USA

We are indebted to Bernd Meyer and Marc Schädler for their versions of Gabor filters that we routinely use. And last but not least, we acknowledge the support of NSF Award 1248047.Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of theNational Science Foundation.

Note: title of the NSF grant at ICSI is Towards Modeling Human Speech Confusions in Noise.
This was a project of the Speech Group.

Introduction

One of our long-term goals is to duplicate in automatic speech recognition (ASR) the robust properties of human speech recognition (e.g., cocktail party effect, word recognition in noise, insensitivity to a wide range of speaking rates, etc.). While it is possible that such significant improvements might be provided without incorporating biological models, making use of known properties of these natural systems could provide a shortcut to better systems, since the design space for signal representations is essentially infinite. Towards this end, we think that it is desirable to improve our models of hearing by incorporating more properties observed in mammals.

The work reported here was limited by available resources (funded only by a modest NSF EAGER grant), so we chose to correspondingly limit our research plan. Rather than incorporate an elaborate auditory model as our baseline, we used a standard front end that is commonly used in ASR, namely, MFCCs, which do model some basic properties of hearing; in a later refinement, we also tried using spectrograms processed with steps from the Power Normalized Cepstral Coefficients (PNCC) approach (Kim and Stern, 2012), which has shown some improvements over PLP or MFCC in some tests on noisy speech. We then applied Gabor filters over a range of temporal modulations, and used the resulting features as inputs to a neural network that was trained to generate features for a Gaussian mixture based HMM ASR system.

We have previously conducted many ASR studies using such Gabor-filtered inputs, which were modeled after spectro-temporal receptive fields that have been measured in experiments by Shamma, Mesgarani, and others (Mesgarani et al, 2008). However, in this case, our effort differed in at least two key aspects. First, our source material was much simpler (CVC syllables, recorded at UCLA), so that we could observe the distribution of accuracies in order to derive correlations with human perception. And secondly, the specific goal of these studies was not to improve ASR, but rather to observe whether these specific modifications of a standard ASR signal representation would improve the correlations with measures of hearing. One intriguing notion to be tested was, at least in this limited example, whether using representations inspired by physiological measurements (the Gabor approach to modulation processing) would provide a better match to psychoacoustics.

Physiological Background for Spectro-Temporal Modulation Filtering

2D Gabor filters appear to closely resemble the spectro-temporal response fields of neurons in the primary auditory cortex, and in particular are used to extract features that simultaneously capture spectral and temporal modulation frequencies for automatic speech applications, as they are used to extract spatial-temporal modulation frequencies for image processing applications (De-Valois and De-Valois, 1990). The overall sensitivity pattern for human hearing has also been observed via perceptual experiments, e.g., from Chi et al (1999) and Drullman et al. (1994). It was observed that humans are most sensitive to temporal modulation frequencies up to 16 Hz and spectral modulation frequencies up to 2 cycles per octave.

A reasonable representation of this process, and one that we have successfully used in speech processing tasks (e.g., speech recognition and speech activity detection), is to filter the time-frequency plane (currently represented by log mel spectral energies extracted every 10 ms from a 25 ms window) with multiple Gabor impulse responses. Each impulse response is complex, with the real component being a cosine windowed by a truncated Gaussian (or a Hann window), and the imaginary being a windowed sine function. Note that such a spectro-temporal impulse response selectively emphasizes particular components of the temporal and spectral modulations that constitute a time-frequency representation. The output of each of these filters (or in some cases of a group of these filters) is further processed by a multi-layer perceptron (MLP) that has been trained to discriminate between phonetic classes. The MLP outputs can then be interpreted as posterior probabilities for these classes, and can be combined additively with weights that are computed dynamically in order to take advantage of the properties of the different feature "streams" for each new acoustic situation. For modeling of human speech perception, the MLP posterior estimates can be used for comparison to experimental phonetic confusions.

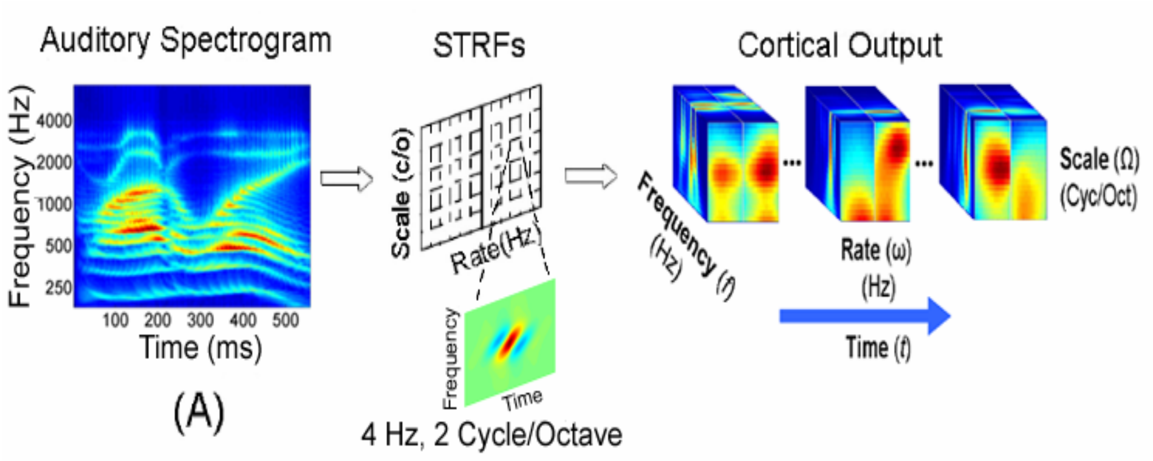

Figure 1: Courtesy of S. Shamma, U. of Maryland (from a presentation). The figure shows the Shamma model of spectro-temporal receptive fields as they have been observed in mammalian primary auditory cortex. Note that "Rate" refers to temporal modulations, "Scale" refers to Spectral modulations, and the Frequency axis is the auditory frequency axis associated with some approximation to the warping of frequency by cochlear function.

Methodology

Collection of the UCLA CVC stimuli

A set of 36 CVC (consonant-vowel-consonant) phonetically balanced syllables was selected that incorporated 13 consonants and 3 vowels. Recordings took place at UCLA using an AKG C-410 head mounted microphone in a soundproof room, with two talkers (one male and one female). Each CVC was repeated twice by each speaker. Babble noise from the Noisex database (Varga and Steeneken, 1993) was added to the CVCs to prepare noisy stimuli. The SNR was calculated by using the average SNR level over the speech-only segment, which was then used to determine the noise power to be added. Each stimulus was prefixed with 100 ms of pure noise (at the noise power calculated in the previous step) in order to enable listeners to adapt to the noise environment. Stimuli were generated corresponding to 6 conditions: 3 SNRs (quiet, 5 dB, 0 dB) x 2 speaking rates (slow, fast). Given the two speakers and the two repetitions, this yielded (36x2x2x2x3) = 864 utterances for both perceptual and ASR experiments.

Conducting Perceptual Experiments

Listening experiments were conducted with 52 subjects, in a soundproof booth at UCLA using the stimuli described above. The subjects would hear the set of 864 CVC stimuli (36 syllables x 2 talkers x 2 speaking rates x 3 noise levels x 2 repetitions) over two sessions of one hour each, corresponding to 432 stimuli per session. The stimuli were played back to back, and the subjects were given a 3 second window between the stimuli in order to respond. They were asked to repeat back the stimulus that they heard. A short break of 10 seconds was given after every 20 stimuli. Two phonetically trained linguists transcribed these manually.

For analysis of the perceptual data, effects were observed for each of the generation factors, but for the purpose of this report, the focus will be on the speaking rate characteristics over the range of consonants.

Automatic Speech Recognition (ASR)

For the ASR, we used a neural network to generate features, and a Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) approach to the modeling of the CVCs. Given the limited amount of CVC data, we primarily used 51 hours of read Wall Street Journal speech (22,092 utterances) for training of both the MLP feature generator and the GMM/HMM acoustic models. However, we adapted the MLP to 471 of the CVC utterances (unused for other testing). Testing was then done on an independent set of 864 CVC utterances. Results for recognition of the different consonants were then compared to the UCLA perceptual results for the different front end models that were considered, and for differing noise levels and speaking rates.

Baseline acoustic front end for ASR experiments

In the 1930s, researchers at places like Bell Labs used filter banks for speech analysis, particularly for experimental analysis-synthesis systems. As speech recognition efforts began in earnest in the 1960s and 70s, filter banks were adapted from earlier applications as a reasonable way to characterize the different speech sounds for classification by machine learning methods of the time. By the mid-late 1980s, it was already customary to use enhancements of these filter banks to emphasize properties of hearing, most notably: (1) some approximation to critical band integration of spectral components (roughly linear at low frequencies and logarithmic at high frequencies, both for bandwidth and spacing); (2) intensity-to-loudness compression of each band's energy; and (3) equal-loudness pre-emphasis (in particular de-weighting low frequencies).

In general, as has become standard for these spectral measures to be transformed into cepstral ones, primarily for the orthogonality property of the latter features, which is desirable for Gaussian modeling. As of this writing, this transformation is sometimes falling out of favor, particularly when it is to be followed by a neural network, which tends to be less "fussy" about preferring decorrelated inputs. However, for our baseline systems, we followed the older convention, since the baseline feature vectors are modeling by a GMM/HMM approach.

Our primary baseline features were Mel Frequency Cepstral Coefficients (MFCCs), which are the most common front ends for GMM/HMM based ASR systems. The filter spacing uses the mel scale (which was derived from pitch perception, and so is not really ideal for spectral estimation, but it has become a standard), and the compression is logarithmic (built into the cepstral transformation). The pre-emphasis typically has a zero at zero frequency as a practical matter, in order to handle the effects of d.c. that can be present in recording equipment.

In addition to twelfth-order MFCCs themselves, we followed establish practice and augmented the 13 (C0-C12) features with their first and second differences (the so-called "delta" features). The features were also normalized to zero mean at the utterance level.

There are many other front end features that, arguably, better represent properties of human hearing, but for the purpose of this study we used this common method as our baseline, primarily to compare with the application of a single combination of methods chosen to a specific simplified model of mammalian hearing.

Acoustic features under test

We augmented the MFCC-based features with features derived from three primary steps: (1) Gammatone-based (Patterson, 1992) spectrogram estimation (2) Gabor filtering (Mesgarani et al, 2008); and (3) neural networks trained as described below. We will refer to this augmentation vector as GT-Gabor-MLP.

We also used a variant of PNCC, which also incorporates gammatone-base spectral estimation, but also includes (1) “medium-time” nonlinear processing that suppresses the effects of additive noise and room reverberation, and (2) a power nonlinearity with exponent 1/15. The final transformation from spectral to cepstral coefficients is performed for PNCC but is skipped or PNS.

Gabor filtering is briefly described in the introduction to this report. For these experiments, seven ranges of temporal modulations, including four in the range of from 0 to 6.2 Hz (centered at 0, 2.4, 3.9, and 6.2 Hz) and three from 9.9 to 25 Hz (centered at 9.9, 15.7, and 25 Hz) were used. For each feature stream, there were 9 spectral modulations ranging from -0.25 to 0.25 cycles per channel leading to a total of 125 inputs (as shown in Table 1 below).

Spectral modulation frequency (cycle/channel)	Number of outputs
0.25, -0.25	40
0.12, -0.12	13
0.06, -0.06	5
0.03, -0.03	3
0	3

Table 1: Spectral modulation ranges used for each temporal modulation frequency

For each of the seven Gabor-filtered PNS-spectrogram inputs, an MLP with a single hidden layer of 1000 neurons was trained to classify one of 41 phonetic units. During testing, the posterior probability outputs were combined using weights derived from the inverse of the entropy of each posterior vector (normalized so that the seven weights added up to one). The weighted and summed outputs were then transformed by a logarithm and PCA dimensionality reduction to yield 25 features, which were then used to augment the MFCC-based baseline features.

GMM/HMM Modeling

As noted above, we trained our GMM/HMM systems with 51 hours of speech from the Wall Street Journal corpus. The acoustic models use cross-word triphones as the modeled units, with statistics estimated with a maximum likelihood procedure. Each triphone is modeled using a 3-state HMM with no skip states. The resulting triphone states are clustered using a decision tree to 5000 tied states. The output distribution for each tied state is modeled with a mixture of 32 multivariate Gaussians with diagonal covariance matrices. The word level model consists of a silence state at the start, followed by an initial /a/ and then a branch to all 36 possible CVCs, concluding with a final silence. HTK was used for both training and decoding.

Determination of results

Following decoding with the word models described above, scoring was done to determine accuracy of the consonant recognition. The vector of consonant accuracies was then compared to UCLA's human perception results via both inspection and correlation.

Results

Here we summarize some of the most prominent results from the study.
The strongest effect observed is shown in Table 2. Note that the 3rd front end method (MFCC + PN-GT-Gabor) only differs from the second by the use of medium-term power normalization, and the use of 1/15 exponent power transformation. Here it appears that, for the case of noisy CVCs, the joint incorporation of these characteristics not only provide a significant improvement in accuracy (although it's still terrible), but also give a better correlation to human perception. However, the use of Gabor modulation filters by themselves (along with a gammatone spectral analysis and MLP transformation of the Gabor output) does not significantly improve accuracy, and appear to even worsen correlation with perception.

Front end method	Machine consonant accuracy in 5dB SNR	Correlation with perception in 5 dB SNR
Baseline (MFCC)	14.9%	0.24
MFCC + GT-Gabor-MLP	15.3%	0.21
MFCC + PN-GT-Gabor-MLP	18.2%	0.28

Table 2: Front end effects for noisy CVCs

For clean speech, on the other hand, neither of the alternative strategies yielded better performance by augmenting MFCCs. It is not clear what we can infer from this, since the additional features incorporate a trained component, and this aspect might have limited the results due to the small amount of training used. See Table 3 for a summary of these results. However, once again it is clear that the additional front end steps are to be preferred (at least they degrade the results less!)

Front end method	Machine consonant accuracy in high SNR	Correlation with perception in high SNR
Baseline (MFCC)	75.7%	0.96
MFCC + GT-Gabor-MLP	67.0%	0.89
MFCC + PN-GT-Gabor-MLP	73.4%	0.93

Table 3: Front end effects for high SNR CVCs

For rapid speech in noise, there is a significant degradation of all accuracies (e.g., from the 14.9% for MFCCs overall to 12.5% for the rapid speech component); however, as shown in Table 4, the correlation with perception for the high modulation Gabor filters actually increases, and the accuracy is far better than what we achieve when all modulation filters are used. Here we are only showing the PN-GT case, since the GT case degrades further in noise (but its results are available in the Appendix). Interestingly, although the accuracy for the low modulations is worse, the correlation with perception is just as strong as for the high modulations.

Front end method	Machine consonant accuracy in 5dB SNR	Correlation with perception in 5 dB SNR
Baseline (MFCC)	12.5%	0.20
MFCC + PN-GT-Gabor-MLP	13.2%	0.26
MFCC + PN-GT-Gabor-MLP low modulations only	10.2%	0.31
MFCC + PN-GT-Gabor-MLP high modulations only	16.5%	0.31

Table 4: Front end effects for noisy CVCs, rapid speech

There is a related effect for slower speech, as illustrated in Table 5; the use of features from low temporal modulations not only yield better accuracy than using MFCCs alone, but also yield better accuracy than just using all of the Gabor filters. However, in this case, some of the results are not as strong. For instance, the correlation number is not better than the baseline. And in this case, using the higher modulations still provide the highest accuracy, and equivalent correlation, despite the test data being slower rate speech.

Front end method	Machine consonant accuracy in 5dB SNR	Correlation with perception in 5 dB SNR
Baseline (MFCC)	17.4%	0.24
MFCC + PN-GT-Gabor-MLP	16.9%	0.19
MFCC + PN-GT-Gabor-MLP low modulations only	19.3%	0.24
MFCC + PN-GT-Gabor-MLP high modulations only	20.0%	0.24

Table 5: Front end effects for noisy CVCs, slower speech

In general, the high modulations tend to do as well or better in all conditions, as highlighted by Tables 6 and 7.

Testing condition	Machine consonant accuracy , MFCC + PN-GT-Gabor-MLP	Machine consonant accuracy, MFCC + PN-GT-Gabor-MLP high modulations only
Clean, all speech	73.4%	74.8%
Clean, rapid speech	71.9%	74.4%
Clean, slow speech	74.9%	75.3%
Noisy, all speech	15.1%	18.2%
Noisy, rapid speech	13.2%	16.5%
Noisy, slow speech	16.9%	20.0%

Table 6: Comparison between using all modulation filters and only the high ones for all testing conditions, using the PN-GT-Gabor-MLP front end, consonant accuracies.

Table 6 shows that the accuracies are always improved (often significantly) by the constraint to higher modulations; while Table 7 shows that the correlations with the results from perceptual tests are similarly improved. There are many possible reasons for this behavior; for instance, the MFCC features to which the modulation features are appended are essentially broad in modulation range, and so emphasizing the higher modulations may help to boost an important region.

We note that for the case of high modulations only, the TOTAL number of trained parameters is significantly lower, since only some of the MLPs are used and the number of hidden units is kept constant.

Testing condition	Machine correlation with perception, MFCC + PN-GT-Gabor-MLP	Machine correlation with perception, MFCC + PN-GT-Gabor-MLP high modulations only
Clean, all speech	0.93	0.95
Clean, rapid speech	0.92	0.94
Clean, slow speech	0.92	0.95
Noisy, all speech	0.22	0.28
Noisy, rapid speech	0.26	0.31
Noisy, slow speech	0.19	0.24

Table 7: Comparison between using all modulation filters and only the high ones for all testing conditions, using the PN-GT-Gabor-MLP front end, correlation with perception.

Full tables of the resulting accuracies are shown in the Appendices to this report.

Discussion

For most of the tests, any strategy that improved consonant accuracy also yielded a better correlation with human performance. However, all of the measured ASR accuracies were far lower than the human perceptual case, particularly for the cases of noise, which is not a surprise. Arguably, having significantly more relevant training data could improve these numbers, but this was not the point of the study. The more relevant question was whether higher temporal modulations were indeed more effective in recognizing faster speech, and this has been demonstrated. However, in case of additive noise, these modulations still provided an improvement, perhaps because the MFCCs covered a range of modulations (and likely were dominated by lower temporal modulations).

In noise, the high temporal modulation additions to the front-end model (using other improvements based on gammatone spectral analysis, power normalization, and 1/5 root compression) did appear to provide better accuracy and correlation with perceptual consonantal accuracies. However, the lack of improvement in the model for the high SNR case (and even an overall reduction in correlation with the perceptual results) indicates that our current approach does not provide an overall effective improvement in modeling. However, we are using far less training data than an adult human has had access to. Perhaps this experiment could be repeated with far more training data, both for the original training data and for the adaptation data. Furthermore, a follow-up experiment should incorporate adaptation of the GMMs, and not only of the MLPs. While this would also be expected to improve the baseline results, our experience with MLPs (and discriminant training of all types) suggests that they could be even more susceptible to overfitting an insufficient amount of training data.

Concluding Remarks

To a speech recognition researcher, the most obvious conclusion is that our system's performance is far worse than what can be observed in human perception.

Still, it is apparent that the use of high temporal modulations is quite effective in improving performance for rapid speech, as one might expect. It is particularly interesting that the reduction of the inputs to filter out low temporal modulations at the input actually improves performance, despite the incorporation of machine learning techniques that in principle would handle the variability in speaking rate.

Acknowledgments

We are indebted to Bernd Meyer and Marc Schädler for their versions of Gabor filters that we routinely use. And last but not least, we acknowledge the support of NSF Award 1248047.

References

Chi, T., Gao, Y., Guyton, M.C., Ru, P., and Shamma, S.A., (1999), “Spectro-temporal modulation transfer functions and speech intelligibility,” *J. Acoust. Soc. Am*., 106(5):719-2732.

Depireux, D.A, Simon, J.Z., Klein, D.J., and Shamma, S.A., (2001), “Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex,“*J. Neurophysiology*, 85:1220-134.

De-Valois, R., and De-Valois, K., (1990), *Spatial Vision*, New York: Oxford University Press.

Drullman, R., Festen, J.M., and Plomp, R. (1994), “Effect of temporal envelope smearing on speech reception,”*J. Acoust. Soc. Am*. 95, 1053-1064.

Hermansky,, H. (1990), “Perceptual linear predictive (PLP) analysis of speech,” *JASA* 87 (4), 1738-1752.

Kim, C., and Stern, R. (2012), “Power-normalized cepstral coefficients (PNCC) for robust speech recognition,” *Proc. ICASSP 2012*, Kyoto, pp. 4101-4104.

Kleinschmidt, M., and Gelbart, D. (2002), “Improving Word Accuracy with Gabor Feature Extraction,” *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, 25-28.

Mesgarani, N., Slaney, M., and Shamma, S.A. (2004), “Speech discrimination based on multiscale spectro-temporal features,” *Proc. ICASSP 2004*, Montreal, 601-604.

Mesgarani, N., David, S.V., Fritz, J.B., and Shamma, S.A. (2008), "Phoneme representation and classification in primary auditory cortex," *J Acoust Soc Am,* vol. 123, Feb 2008, 899-909.

Mesgarani, N., Shamma, S.A. (2011), “Speech processing with a cortical representation of audio,” *Proc. ICASSP 2011*, Prague, 5872-5875.

Meyer, B.T., Ravuri, S.V., Schaedler, M.R., and Morgan, N. (2011), “Comparing Different Flavors of Spectro-Temporal Features for ASR,” *Proc. Interspeech 2011*, Florence, Italy, 1269-1272.

Patterson, R.D. (1976). “Auditory filter shapes derived by noise stimuli,” *J. Acoust. Soc. Am*. 59, 640–654.

Varga, A., and Steeneken, H. (1993), “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, Issue 3, July 1993, pp. 247-251.

Zhao, S.Y., Ravuri, S., and N. Morgan, N. (2009), *“*Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR,” *Proc. Interspeech 2009,* 2951-2954.

Appendix A: Full results for PN-GT-Gabor-MLP augmentation of MFCCs

Table A1 – High SNR results, consonant accuracy, speech overall

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	75.7%	0.96
PN-Gabor-MLP	73.4%	0.93
PN-Gabor-MLP-low mod	69.2%	0.86
PN-Gabor-MLP-high mod	74.8%	0.95

Table A2 – High SNR results, consonant accuracy, rapid speech

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	76.7%	0.94
PN-Gabor-MLP	71.9%	0.92
PN-Gabor-MLP-low mod	65.5%	0.86
PN-Gabor-MLP-high mod	74.4%	0.94

Table A3 – High SNR results, consonant accuracy, slower speech

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	74.7%	0.94
PN-Gabor-MLP	74.9%	0.92
PN-Gabor-MLP-low mod	72.9%	0.86
PN-Gabor-MLP-high mod	75.3%	0.95

Table A4 – 5 dB SNR results, consonant accuracy, speech overall

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	14.9%	0.24
PN-Gabor-MLP	15.1%	0.22
PN-Gabor-MLP-low mod	14.7%	0.26
PN-Gabor-MLP-high mod	18.2%	0.28

Table A5 – 5 dB SNR results, consonant accuracy, rapid speech

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	12.5%	0.20
PN-Gabor-MLP	13.2%	0.26
PN-Gabor-MLP-low mod	10.2%	0.31
PN-Gabor-MLP-high mod	16.5%	0.31

Table A6 – 5 dB SNR results, consonant accuracy, slower speech

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	17.4%	0.24
PN-Gabor-MLP	16.9%	0.19
PN-Gabor-MLP-low mod	19.3%	0.22
PN-Gabor-MLP-high mod	20.0%	0.24

Appendix B: Full results for GT-Gabor-MLP augmentation of MFCCs

Table B1 – High SNR results, consonant accuracy, speech overall

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	75.7%	0.96
GT-Gabor-MLP	67.0%	0.89
GT-Gabor-MLP-low mod	57.2%	0.78
GT-Gabor-MLP-high mod	68.7%	0.87

Table B2 – High SNR results, consonant accuracy, rapid speech

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	76.7%	0.94
GT-Gabor-MLP	65.1%	0.87
GT-Gabor-MLP-low mod	51.1%	0.68
GT-Gabor-MLP-high mod	70.3%	0.88

Table B3 – High SNR results, consonant accuracy, slower speech

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	74.7%	0.94
GT-Gabor-MLP	65.1%	0.87
GT-Gabor-MLP-low mod	63.3%	0.81
GT-Gabor-MLP-high mod	67.2%	0.81

Table B4 – 5 dB SNR results, consonant accuracy, speech overall

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	14.9%	0.24
GT-Gabor-MLP	15.3%	0.21
GT-Gabor-MLP-low mod	11.1%	0.15
GT-Gabor-MLP-high mod	6.5%	0.09

Table B5 – 5 dB SNR results, consonant accuracy, rapid speech

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	12.5%	0.20
GT-Gabor-MLP	13.8%	0.27
GT-Gabor-MLP-low mod	10.3%	0.21
GT-Gabor-MLP-high mod	4.3%	0.10

Table B6 – 5 dB SNR results, consonant accuracy, slower speech

Front end method	Machine accuracy	Correlation with perception
Baseline (MFCC)	17.4%	0.24
GT-Gabor-MLP	16.8%	0.18
GT-Gabor-MLP-low mod	11.9%	0.11
GT-Gabor-MLP-high mod	9.0%	0.10