

# IMPROVING WORD ACCURACY WITH GABOR FEATURE EXTRACTION

Michael Kleinschmidt <sup>a,b</sup> and David Gelbart <sup>a</sup>

<sup>a</sup> International Computer Science Institute Berkeley, CA, USA

<sup>b</sup> Medizinische Physik, Universität Oldenburg, Germany  
{michaelk, gelbart}@ICSI.berkeley.EDU

## ABSTRACT

A novel type of feature extraction for automatic speech recognition is investigated. Two-dimensional Gabor functions, with varying extents and tuned to different rates and directions of spectro-temporal modulation, are applied as filters to a spectro-temporal representation provided by mel spectra. The use of these functions is motivated by findings in neurophysiology and psychoacoustics. Data-driven parameter selection was used to obtain Gabor feature sets, the performance of which is evaluated on the Aurora 2 and 3 datasets both on their own and in combination with the Qualcomm-OGI-ICSI Aurora proposal. The Gabor features consistently provide performance improvements.

## 1. INTRODUCTION

Speech is characterized by its fluctuations across time and frequency. The latter reflect the characteristics of the human vocal cords and tract and are commonly exploited in automatic speech recognition (ASR) by using short-term spectral representations such as cepstral coefficients. The temporal properties of speech are targeted in ASR by dynamic (delta and delta-delta) features and temporal filtering and feature extraction techniques like RASTA and TRAPS [1]. Nevertheless, speech clearly exhibits combined *spectro-temporal* modulations. This is due to intonation, coarticulation and the succession of several phonetic elements, e.g., in a syllable. Formant transitions, for example, result in diagonal features in a spectrogram representation of speech. This kind of pattern is explicitly targeted by the feature extraction method used in this paper.

Recent findings from a number of physiological experiments in different mammal species showed that a large percentage of neurons in the primary auditory cortex respond differently to upward- versus downward-moving ripples in the spectrogram of the input [2]. Each individual neuron is tuned to a specific combination of spectral and temporal modulation frequencies, with a spectro-temporal response

field that may span up to a few 100ms in time and several critical bands in frequency and may have multiple peaks [3, 4]. A psychoacoustical model of modulation perception [5] was built based on that observation and inspired the use of two-dimensional Gabor functions as a feature extraction method for ASR in this study. Gabor functions are localized sinusoids known to model the characteristics of neurons in the visual system [6]. The use of Gabor features for ASR has been proposed earlier and proven to be relatively robust in combination with a simple classifier [7]. Automatic feature selection methods are described in [8] and the resulting parameter distribution has been shown to remarkably resemble neurophysiological and psychoacoustical data as well as modulation properties of speech. Other approaches to targeting spectro-temporal variability in feature extraction include time-frequency filtering (tiffing) [9]. Still, this novel approach of spectro-temporal processing by using localized sinusoids most closely matches the neurobiological data and also incorporates other features as special cases: purely spectral Gabor functions perform sub-band cepstral analysis—modulo the windowing function—and purely temporal ones can resemble TRAPS or the RASTA impulse response and its derivatives [1] in terms of temporal extent and filter shape.

## 2. SPECTRO-TEMPORAL FEATURE EXTRACTION

A spectro-temporal representation of the input signal is processed by a number of Gabor functions used as 2-D filters. The filtering is performed by correlation over time of each input frequency channel with the corresponding part of the Gabor function (with the Gabor function centered on the current frame and desired frequency channel) and a subsequent summation over frequency. This yields one output value per frame per Gabor function (we call these output values the Gabor features) and is equivalent to a 2-D correlation of the input representation with the complete filter function and a subsequent selection of the desired frequency channel of the output.

In this study, log mel-spectrograms serve as input fea-

---

This work was supported by Deutsche Forschungsgemeinschaft (KO 942/15), the Natural Sciences and Engineering Research Council of Canada, and the German Ministry for Education and Research.

tures for Gabor feature extraction. This was chosen for its widespread use in ASR and because the logarithmic compression and mel-frequency scale might be considered a very simple model of peripheral auditory processing. Any other spectro-temporal representation of speech could be used instead and especially more sophisticated auditory models might be a good choice for future experiments.

The two-dimensional complex Gabor function  $g(t, f)$  is defined as the product of a Gaussian envelope  $n(t, f)$  and the complex Euler function  $e(t, f)$ . The envelope width is defined by standard deviation values  $\sigma_f$  and  $\sigma_t$ , while the periodicity is defined by the radian frequencies  $\omega_f$  and  $\omega_t$  with  $f$  and  $t$  denoting the frequency and time axis, respectively. The two independent parameters  $\omega_f$  and  $\omega_t$  allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including *diagonal* modulations. Further parameters are the centers of mass of the envelope in time and frequency  $t_0$  and  $f_0$ . In this notation the Gaussian envelope  $n(t, f)$  is defined as

$$n(\cdot) = \frac{1}{2\pi\sigma_f\sigma_t} \cdot \exp \left[ \frac{-(f - f_0)^2}{2\sigma_f^2} + \frac{-(t - t_0)^2}{2\sigma_t^2} \right] \quad (1)$$

and the complex Euler function  $e(t, f)$  as

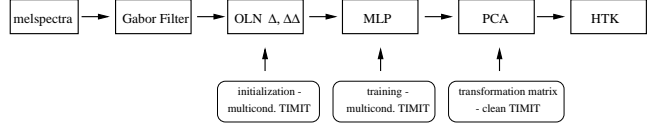
$$e(\cdot) = \exp [i\omega_f(f - f_0) + i\omega_t(t - t_0)]. \quad (2)$$

It is reasonable to set the envelope width depending on the modulation frequencies  $\omega_f$  and  $\omega_t$  to keep the same number of periods  $T$  in the filter function for all frequencies. Here, the spread of the Gaussian envelope in dimension  $x$  was set to  $\sigma_x = \frac{\pi}{\omega_x} = T_x/2$ . The infinite support of the Gaussian envelope is cut off at between  $\sigma_x$  and  $2\sigma_x$  from the center. For time dependent features,  $t_0$  is set to the current frame, leaving  $f_0$ ,  $\omega_f$  and  $\omega_t$  as free parameters. From the complex results of the filter operation, real-valued features may be obtained by using the real or imaginary part only. In this case, the overall DC bias was removed from the template. The magnitude of the complex output can also be used. Special cases are temporal filters ( $\omega_f = 0$ ) and spectral filters ( $\omega_t = 0$ ). In these cases,  $\sigma_x$  replaces  $\omega_x = 0$  as a free parameter, denoting the extent of the filter, perpendicular to its direction of modulation.

### 3. ASR EXPERIMENTS

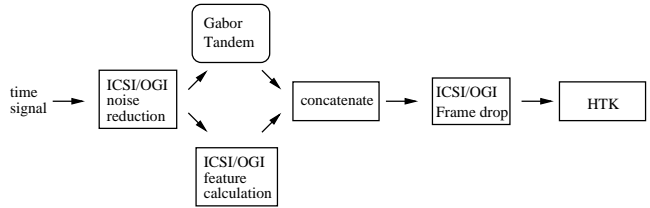
#### 3.1. Set up

The Gabor features approach is evaluated within the aurora experimental framework [10] using a) the Tandem recognition system [11] and d) a combination of it with the Qualcomm-ICSI-OGI QIO-NoTRAPS system, which is described in [12]. Variants of that are b) and c): the Gabor Tandem system as a single stream combined with noise robustness techniques taken from the Qualcomm-ICSI-OGI proposal.



**Fig. 1.** Sketch of the Gabor Tandem recognition system as it was used in experiment a).

In all cases the Gabor features are derived from log mel-spectrograms, calculated as in [13] but modified to output mel-spectra instead of MFCCs, omitting the final DCT. The log mel-spectrogram calculation consists of DC removal, pre-emphasis, Hanning windowing with 10ms offset and 25ms length, FFT and summation of the magnitude values into 23 mel-frequency channels with center frequencies from 124 to 3657Hz. The amplitude values are then compressed by the natural logarithm.



**Fig. 2.** Experiment d): Combination of Gabor feature extraction and the Qualcomm-ICSI-OGI proposal system.

Fig. 1 sketches the Tandem system as it is used in experiment a): 60 Gabor filters are fed into a multi-layer perceptron (MLP) after online normalization (OLN) and  $\Delta, \Delta\Delta$  processing. The MLP (180 input, 1000 hidden, 56 output units) has been trained on the frame labeled noisy TIMIT corpus using frame by frame phoneme targets. The output layer's softmax non-linearity is omitted in forward passing. The resulting 56-dimensional feature vector is then decorrelated by a PCA transform based on clean TIMIT. The resulting feature vectors are then given to the fixed Aurora HTK back end.

Experiment d) is depicted in Fig. 2. After the initial noise reduction (NR), which is the same as in [12], a Gabor feature stream identical to that in a) is run in parallel with the Qualcomm-ICSI-OGI proposal feature extraction. The two streams are combined by concatenation before the final frame dropping (FD) of frames judged to be nonspeech. The 45 Qualcomm-ICSI-OGI features are combined with a reduced set of 15 features from the Gabor stream which are obtained by reducing the dimensionality in the PCA stage from 56 to 15. In a variation of this, experiment c), the full set of 56 features from the Gabor stream is used with noise reduction and frame dropping but without concatenating the Qualcomm-ICSI-OGI feature stream. Experiment

Aurora 2	WER [%]		Rel. impr. [%]	
	multi	clean	multi	clean
R0: Aurora2 reference	12.97	41.94	0.00	0.00
R1: ICSI/OGI	9.09	15.10	26.41	66.53
R2a) T melspec	12.04	28.66	12.87	40.09
R2d): R1 + T melspec NR FD	9.18	14.01	34.55	72.29
G1a) T Gabor	11.68	30.17	14.52	37.19
G2a) T Gabor	11.99	26.51	8.40	44.42
G3a) T Gabor	11.99	23.63	4.03	51.24
G1b) T Gabor NR	10.33	16.51	19.88	64.64
G1c) T Gabor NR FD	10.42	14.42	25.74	70.86
G1d) R1 + T Gabor NR FD	8.85	13.04	37.84	74.99
G2d) R1 + T Gabor NR FD	8.70	13.30	37.65	73.88
G3d) R1 + T Gabor NR FD	8.60	12.29	36.40	75.23

**Table 1.** Aurora 2 (TIDigits): Performance of different front ends in terms of WER and WER reduction relative to the baseline system (R0). The Qualcomm-ICSI-OGI submission system (R1) is compared and combined with different Gabor Tandem (T) systems: Gabor set G1 was optimized on TIMIT phoneme inter-group discrimination, G2 on TIMIT phoneme inter- and within-group discrimination and G3 on German digits. NR indicates noise reduction, FD frame dropping. R2 denotes a Tandem system based on mel spectra.

b) also leaves out the frame dropping stage.

Reference systems are the aurora baseline (R0) front end of 13 mel-cepstral coefficients and their delta and double-deltas used in the unquantized, endpointed version [14], the Qualcomm-ICSI-OGI proposal system (R1), and a combination of R1 with a melspec-based Tandem system (R2) which is identical to the Gabor-based Tandem system used apart from the input features to the MLP, which are 23 melspectra with deltas and double deltas over 90ms (9 frames) of context. Also, the number of hidden units has been reduced to 300 in order to keep the total number of weights constant.

In the Aurora 2 experiment, training and testing use the TIDigits English connected digits corpus, artificially mixed with noise of varying levels and types. HTK is trained separately with clean and multi-condition training data. Test set A refers to matched noise (in the case of multicondition training), test set B to mismatched noise and test set C to mismatched channel conditions. For Aurora 3 training and testing use the Speechdat-car corpora for Finnish, Spanish, German and Danish [14]. The corpora contain digits strings recorded in various car environments. The experimental results refer to well-matched (wm), medium-mismatched (mm) and highly-mismatched (hm) conditions which describe the degree of mismatch of noise and microphone location (close-talking versus hands-free) between the training and test sets. mm indicates a mismatch in noise only, while hm indicates mismatch of noise and microphone.

	Aurora 2		Aurora 3		overall	
	WER [%]	impr. [%]	WER [%]	impr. [%]	WER [%]	impr. [%]
R0	27.46	0.00	23.48	0.00	25.47	0.00
R1	12.10	46.47	9.43	53.94	10.77	50.21
R2 d)	11.60	53.42	9.23	56.73	10.42	55.08
G1 d)	10.95	56.41	9.20	57.60	10.08	57.01
G2 d)	11.00	55.77	8.91	58.28	9.96	57.03
G3 d)	10.44	55.82	8.88	57.44	9.66	56.63

**Table 2.** Aurora2 (TIDigits) and Aurora 3 (speechdat-car): Performance of different front ends in terms of WER and WER reduction. Abbreviations as in Table 1.

### 3.2. Feature selection

The parameters of the 60 Gabor filters were chosen by optimization as described in [7, 8]. A simple linear classifier was used to evaluate the importance of individual feature based on their contribution to classification performance. Gabor set G1 is optimized on inter-group discrimination of phoneme targets from the TIMIT corpus combined into broader phonetic categories of place and manner of articulation. Gabor set G2 is optimized on inter- and within-group discrimination of broad phonetic classes, also using the TIMIT corpus. G3 is optimized on German digits (zifkom corpus) using word targets. G1, G2 and G3 respectively contain 27, 28, and 48 filters with temporal extents longer than 100 ms, although many in G1 are much shorter. Set G1 consists of 35 features with purely spectral modulation, 23 with purely temporal modulation, and two with spectro-temporal modulation. G2 (34/22/4) and G3 (12/18/30) have a larger number of filters with spectro-temporal modulation. In all three cases, most of the features are two-dimensional in extent, simultaneously occupying more than one frequency channel and time frame. Lists of the filter parameters are available online [15].

### 3.3. Results

The results in Tables 1–4 are given in absolute word error rate (WER=1-Accuracy) and WER improvement relative to the baseline system (R0). The WER as well as the WER reduction values are averaged over a number of different test conditions in accordance with [14], so the average WER improvement cannot directly be calculated from the average WERs.

All systems in configuration a) yield better results on the Aurora 2 task than the reference system R0 (cf. Table 1). The three Gabor sets vary in their performance for clean and noisy training conditions. The more spectro-temporal features in the set, the better the performance with clean training, indicating an improved robustness with these features. Adding the NR in b) and the FD in c) further improves the performance.

Aurora 2 Word Error Rate [%]				
	Set A	Set B	Set C	Overall
<b>Multi</b>	8.09	8.77	9.29	<b>8.60</b>
<b>Clean</b>	11.72	13.13	11.74	<b>12.29</b>
<b>Average</b>	<b>9.90</b>	<b>10.95</b>	<b>10.51</b>	<b>10.44</b>

Aurora 2 Relative Improvement [%]				
	Set A	Set B	Set C	Overall
<b>Multi</b>	33.85	37.05	40.23	<b>36.40</b>
<b>Clean</b>	74.94	76.96	72.32	<b>75.23</b>
<b>Average</b>	<b>54.40</b>	<b>57.00</b>	<b>56.27</b>	<b>55.82</b>

**Table 3.** Aurora 2 (TIDigits) WER and relative improvement for system G3d, a combination of the Qualcomm-ICSI-OGI system (R1) and the Gabor Tandem G3 NR FD stream.

Aurora 3 Word Error Rate [%]					
	Finnish	Spanish	German	Danish	Average
<b>wm</b>	2.73	2.14	5.43	6.41	<b>4.18</b>
<b>mm</b>	10.81	4.14	11.71	19.01	<b>11.42</b>
<b>hm</b>	10.25	8.18	11.61	21.39	<b>12.86</b>
<b>all</b>	<b>7.44</b>	<b>4.35</b>	<b>9.17</b>	<b>14.57</b>	<b>8.88</b>

Aurora 3 Relative Improvement [%]					
	Finnish	Spanish	German	Danish	Average
<b>wm</b>	62.40	69.69	38.30	49.61	<b>55.00</b>
<b>mm</b>	44.54	75.19	38.24	41.83	<b>49.95</b>
<b>hm</b>	82.76	83.12	56.73	64.72	<b>71.83</b>
<b>all</b>	<b>61.24</b>	<b>74.97</b>	<b>42.88</b>	<b>50.66</b>	<b>57.44</b>

**Table 4.** Aurora 3 (Speechdat-car) WER and relative improvement for system G3d).

Our best results are obtained by combining R1 with one of the Tandem streams via concatenation in experiment d). Table 2 summarizes the results for Aurora 2 and 3. Combining the Qualcomm-ICSI-OGI feature set (R1) with Tandem based features improves performance on Aurora 2 and 3 in terms of average WER and average WER improvement. Gabor based Tandem systems perform better than the mel spectrum based Tandem system (R2d). System G2d yields the greatest (57.03%) overall relative improvement over R0, while system G3d yields the lowest overall WER (9.66%). This is due to G3 being more robust in very adverse conditions, where the absolute gain in WER is higher. Tables 3 and 4 give more detailed results for feature set G3d).

#### 4. CONCLUSION

Optimized sets of Gabor features have been shown to improve robustness when used as part of the Tandem system. When incorporating the Tandem system as a second stream into the already robust Qualcomm-ICSI-OGI proposal, the overall performance can be increased further by almost 7% absolute in relative WER improvement or over 1% abso-

lute reduction in WER. The fact that Gabor-based Tandem systems consistently outperformed mel spectrum-based systems shows the usefulness of explicitly targeting extended spectro-temporal patterns. In adverse conditions, the Gabor set G3 with 50% diagonal features performs best, which further supports the approach of spectro-temporal modulation filters. It is to be investigated whether this holds for large vocabulary tasks.

*Special thanks go to Barry Yue Chen, Stéphan Dupont, Steven Greenberg, Hynek Hermansky, Birger Kollmeier, Nelson Morgan, and Sunil Sivasdas for technical support and great advice.*

#### 5. REFERENCES

- [1] H. Hermansky, "Should recognizers have ears?," *Speech Communication*, vol. 25, pp. 3–24, 1998.
- [2] D.A. Depireux, J.Z. Simon, D.J. Klein, and S.A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 85, pp. 1220–1234, 2001.
- [3] C.E. Schreiner, H.L. Read, and M.L. Sutter, "Modular organization of frequency integration in primary auditory cortex," *Annu. Rev. Neurosc.*, vol. 23, pp. 501–529, 2000.
- [4] R. C. deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, pp. 1439–1443, 1998.
- [5] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2719–2732, 1999.
- [6] R De-Valois and K. De-Valois, *Spatial Vision*, Oxford U.P., New York, 1990.
- [7] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in ASR," *Acustica united with acta acustica*, 2002, (accepted).
- [8] M. Kleinschmidt, "Spectro-temporal Gabor features as a front end for ASR," in *Proc. Forum Acusticum Sevilla*, 2002.
- [9] C. Nadeu, D. Macho, and J. Hernando, "Time & frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, no. 1–2, pp. 93–144, 2000.
- [10] H.G. Hirsch and D. Pearce, "The Aurora experimental framework ...," in *ISCA ITRW ASR: Challenges for the Next Millennium, Paris*, 2000.
- [11] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP, Istanbul*, 2000.
- [12] A. Adami et al., "Qualcomm-ICSI-OGI features for ASR," in *Proc. ICSLP*, 2002, (submitted).
- [13] "ETSI Standard: ETSI ES 201 108 V1.1.2 (2000-04)," 2000.
- [14] "Aurora," at [icslp2002.colorado.edu/special\\_sessions/aurora](http://icslp2002.colorado.edu/special_sessions/aurora).
- [15] "Gabor feature extraction," at [www.icsi.berkeley.edu/Speech/papers/icslp02-gabor](http://www.icsi.berkeley.edu/Speech/papers/icslp02-gabor).