

# Localized spectro-temporal features for automatic speech recognition

Michael Kleinschmidt

Medical Physics Section, Carl von Ossietzky Universität Oldenburg, Germany

Michael.Kleinschmidt@uni-oldenburg.de

## Abstract

Recent results from physiological and psychoacoustic studies indicate that spectrally and temporally localized time-frequency envelope patterns form a relevant basis of auditory perception. This motivates new approaches to feature extraction for automatic speech recognition (ASR) which utilize two-dimensional spectro-temporal modulation filters. The paper provides a motivation and a brief overview on the work related to Localized Spectro-Temporal Features (LSTF). It further focuses on the Gabor feature approach, where a feature selection scheme is applied to automatically obtain a suitable set of Gabor-type features for a given task. The optimized feature sets are examined in ASR experiments with respect to robustness and their statistical properties are analyzed.

## 1. Getting auditory ... again?

The question whether knowledge about the (human) auditory system provides valuable contributions to the design of ASR systems is as old as the field itself. The topic has been discussed extensively elsewhere (e.g. [1]). After all these years, a major argument still holds, namely the large gap in performance between normal-hearing native listeners and state-of-the-art ASR systems. Consistently, humans outperform machines by at least an order of magnitude [2]. Human listeners recognize speech even in very adverse acoustical environments with strong reverberation and interfering sound sources. However, this discrepancy between human and machine performance is not restricted to robustness alone. It is observed also in undisturbed conditions and very small context independent corpora, where higher level constraints (cognitive aspects, language model) do not play a role. Arguably this hints towards an insufficient feature extraction in machine recognition systems. It is argued here, that including LSTF streams provides another step towards human-like speech recognition.

## 2. Evidence for (spectro-)temporal processing in the auditory system

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 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 [3] and TRAPS [4]. Nevertheless, speech clearly exhibits combined *spectro-temporal* modulations. This is due to intonation, co-articulation 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 captured by LSTF and explicitly targeted by the Gabor feature

extraction method described below.

### 2.1. Neurophysiology

Recent findings from a number of physiological experiments in different mammalian species have revealed the spectro-temporal receptive fields (STRF) of neurons in the primary auditory cortex. Individual neurons are sensitive to specific spectro-temporal patterns in the incoming sound signal. The results were obtained using reverse correlation techniques with complex spectro-temporal stimuli such as checkerboard noise or moving ripples [5]. The STRFs often clearly exceed one critical band in frequency, have multiple peaks and also show tuning to temporal modulation. In many cases the neurons are sensitive to the direction of spectro-temporal patterns (e.g. upward or downward moving ripples, c.f. Fig. 1), which indicates a combined spectro-temporal processing rather than consecutive stages of spectral and temporal filtering [6]. Still the STRF are mainly localized in time and frequency, generally spanning at most 250 ms and one or two octaves, respectively. The center frequency distributions of the linear modulation filter transfer function associated with the STRFs show a broad peak between 4 and 8 Hz in the ferret and at about 12 Hz in the cat [7].

In the visual cortex, STRFs are measured with (moving) orientated grating stimuli. The results very well match two-dimensional Gabor functions [8]. Often, two neurons show very similar STRFs differing only by a  $\pi/2$  phase shift. Two such cells combined provide for a translation-invariant detection of a given modulation pattern within a certain part of the visual field.

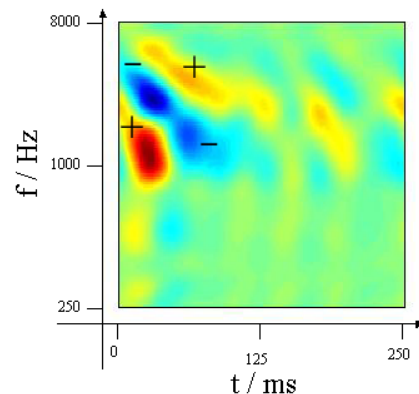


Figure 1: Example of a diagonal STRF from a neuron in the primary auditory cortex of a ferret. Courtesy of David J. Klein [6]. Red colour(+) denotes excitatory, blue(-) inhibitory regions in the receptive field.

## 2.2. Psychoacoustics and human speech perception

The neurophysiological findings fit well with psychoacoustic experiments on early auditory features [9]. A psychophysical reverse correlation technique was applied to analyze subjects performance in masking experiments with semi-periodic white noise. The resulting basic auditory feature patterns are distributed in time and frequency and in some cases comprised of several unconnected parts, very much resembling the STRF of cortical neurons.

In psychoacoustic modelling, temporal modulation filter-bank approaches become more and more accepted. The perception model (PEMO) of effective processing, for example, utilizes a bank of modulation bandpass filters for each single critical band to account for a number of fundamental psychoacoustic experiments [10]. However, psychoacoustical phenomena such as comodulation masking release (CMR) indicate cross-channel mechanisms. More recent models implicitly include localized spectro-temporal filters [11].

When Fletcher et al. examined speech intelligibility of human listeners, they found log sub-band classification error probability to be additive for nonsense syllable recognition tasks. This suggests independent temporal processing in a number of articulatory bands. Their work resulted in the definition of the articulation index, a model of human speech perception [12]. However, recent speech intelligibility experiments have shown that the combination of two distant narrow spectral channels or slits leads to a gain in intelligibility which is greater than predicted by the articulation index (e.g. [13]). The new data suggests some integration of information across frequency bands. Other experiments with artificially distorted modulation amplitude and phase in these bands showed a relatively high tolerance of e.g. phase distortions in speech intelligibility measurements [14]. This would indicate at least partly channel-independent processing. The peripheral channels in these experiments were more than one octave apart, ruling out only global spectral integration of information and still allowing for *localized* spectro-temporal features.

## 3. A brief history of temporal processing in automatic speech recognition

Standard front ends, such as mel-cepstra or perceptual linear prediction, only represent the spectrum within short analysis frames and thereby tend to neglect very important dynamic patterns in the speech signal. This deficiency has been partly overcome by adding temporal derivatives in the form of delta and delta-delta features to the set. Delta features effectively provide for a comb filtering effect in the temporal modulation frequency domain. A number of different modulation filtering techniques in the cepstral or spectral domain have been developed since then. Depending on optional log amplitude compression, channel effects or additive noise can be reduced by temporal bandpass and highpass envelope filtering such as cepstral mean subtraction, RASTA processing [3], the modulation spectrogram [15] or adaptation loops of PEMO processing [16]. The usefulness of modulation bandpass filtering for ASR has been studied in detail and well matches the importance of individual modulation frequency ranges for speech intelligibility for human listeners [17].

New methods of purely temporal processing have been established motivated by Fletcher's findings of independent processing in each frequency channel and the focus on dynamic aspects of speech. Most prominent examples are the TempoRAI

PatternS (TRAPS) [4] which apply multi-layer perceptrons to classify current phonemes in each single critical band based on a temporal context of up to 1 s. Another approach is multi-band processing, for which features are calculated in broader sub-bands to reduce the effect of band-limited noise on the overall performance. However, all these feature extraction methods apply either spectral or temporal processing at a time.

## 4. A trend towards localized spectro-temporal features

Neurophysiology and psychoacoustic research yields auditory features of varying extent and shape which can be categorized as purely spectral, purely temporal or spectro-temporal. In the ASR domain, feature extraction methods have been dominant so far, that are one-dimensional and of large extent in that dimension (such as a cepstral analysis over the whole spectrum at one point in time). Following the biological blueprint and adding localized, 2D spectro-temporal features yields several advantages:

**Diagonality** - LSTF very efficiently detect diagonal structures in spectro-temporal representations of speech such as formant transitions.

**Locality** - the limited extent of LSTF fosters robustness in additive noise and mitigates channels effects.

**Adaptivity** - the size of each individual LSTF can be matched to the type of pattern it is designed for (in contrast to e.g. high cepstral coefficients which are calculated over the whole spectrum for detecting two neighboring spectral peaks).

**Generality** - purely spectral and purely temporal LSTFs are akin to cepstral analysis and modulation bandpass filtering, respectively; therefore the class of LSTFs also includes existing types of feature extraction, with the restriction to localized processing.

There are a number of different approaches to achieve spectro-temporal feature extraction for ASR, such as spectro-temporal modulation filtering [18], linear transformations to the spectrogram representation of speech, e.g. linear discriminant analysis (LDA), independent component analysis (ICA), and principal component analysis (PCA) [19], and the extension of TRAPS to more than one critical band [20]. Approaches to use artificial neural networks for ASR classify spectral features using temporal context on the order of 10 to 100 ms. Depending on the system, this is part of the back end as in the connectionist approach [21] or part of the feature extraction as in the Tandem system [22].

The main problem of LSTF is the large number of possible parameter combinations. This issue may be solved implicitly by automatic learning in neural networks with a spectrogram input and a long time window of e.g. 1 s. However, this is computationally expensive and prone to overfitting, as it requires large amounts of (labeled) training data, which are often unavailable. By putting further constraints on the spectro-temporal patterns, the number of free parameters can be decreased by several orders of magnitude. This is the case when a specific analytical function, such as sigma-pi cells [23] or the Gabor function [24], is explicitly demanded. This approach narrows the search to a certain sub-set and thereby some important features might be ignored. However, neurophysiological and psychoacoustic knowledge can be exploited for the choice of the prototype. Another promising approach is to apply un-supervised

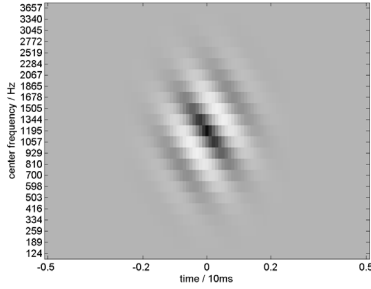


Figure 2: Real part of a discrete two-dimensional complex Gabor function with parameters  $\omega_t/2\pi = 6$  Hz and  $\omega_f/2\pi \approx 0.55$  cycl./oct. centered at  $f_0 = 1195$  Hz. Sampling as for a mel-spectrogram primary feature matrix.

machine learning techniques to derive a suitable set of features. Thus, a sparseness criterion was applied to derive optimal feature from spectro-temporal speech data [25]. The resulting features showed a close resemblance to the STRFs of cortical neurons in the auditory system. Similar techniques are widely used in the visual domain.

## 5. Localized spectro-temporal Gabor features for automatic speech recognition

The STRF of cortical neurons and early auditory features derived in psychoacoustic experiments can be approximated, although somewhat simplified, by two-dimensional Gabor functions. The Gabor filter functions generally target two-dimensional envelope fluctuations but include, as special cases, purely spectral (local cepstra – modulo the windowing function) and purely temporal (modulation bandpass) features. The latter resemble TRAPS or the RASTA impulse response and its derivatives [1] in terms of temporal extent and filter shape. The use of Gabor features for ASR has been proposed earlier and proven to be relatively robust in combination with a simple classifier [24]. By applying a ‘wrapper approach’ to feature selection, optimized sets of Gabor features were obtained that also allowed for increased robustness in adverse acoustic conditions for digit strings in the Aurora 2 & 3 experimental setup [26]. These results were obtained by combining Gabor feature streams with other conventional feature streams.

### 5.1. Background

The two-dimensional complex Gabor function  $g(t, f)$  is defined as the product of a Gaussian envelope  $n(t, f)$  and a complex exponential function  $e(t, f)$ . The envelope width is given 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 exponential  $e(t, f)$  as

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

In order to keep the same number of periods  $T$  in the filter function for all frequencies, the envelope width is set depending on the modulation frequencies  $\omega_f$  and  $\omega_t$ . This essentially yields a 2D-wavelet prototype with the scale factors  $\sigma_t$  and  $\sigma_f$ . Typically, the spread of the Gaussian envelope in dimension  $x$  is set to  $\sigma_x = \frac{\pi}{\omega_x} = T_x/2$ . For time-dependent features,  $t_0$  is set to the current frame, leaving  $f_0$ ,  $\omega_f$  and  $\omega_t$  as free parameters. 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.

Gabor features are derived from a two-dimensional input pattern, typically a series of feature vectors. A number of processing schemes may be considered for these primary features that extract a spectro-temporal representation from the input wave form. The range is from a spectrogram to sophisticated auditory models. From the complex results of the filter operation, real valued features may be obtained by using the magnitude of the complex filter output. This allows for a phase independent feature extraction which closely corresponds to the properties of certain cells in the visual cortex.

Due to the large number of possible parameter combinations, it is necessary to select a suitable set of features. This may be carried out by a modified version of the Feature-finding Neural Network (FFNN). It consists of a linear single-layer perceptron in conjunction with secondary feature extraction and an optimization rule for the feature set [27]. The linear classifier guarantees fast training. This is mandatory because, with the applied wrapper method of feature selection, the importance of each feature is evaluated by the increase of RMS classification error after its removal from the set, requiring iterative re-training of the classifier.

### 5.2. Experimental results

If the primary feature matrix consists of log mel-spectrum calculated as in the ETSI aurora reference (ETSI ES 201 v1.1.2), the range of parameters is limited mainly by the resolution of the primary feature matrix (100 Hz and 23 channels covering 6 octaves). Therefore, the temporal modulation frequency is limited to a range of 2 - 50 Hz. In this case, the automated feature selection algorithm described above yields a distribution of parameters that reflects the statistical properties of the training speech material and is well in the range of temporal modulation characteristics of neurophysiological STRFs. Histograms for the temporal modulation frequencies of the 2D filters are given in Fig. 3.

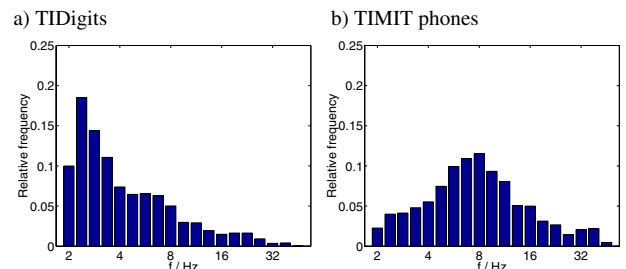


Figure 3: Distribution of temporal modulation frequency  $\omega_t/2\pi$  after automatic feature selection for a) clean TIDigits single word targets (40 sets, 2400 features) and b) clean TIMIT phone targets (146 sets, 2923 features).

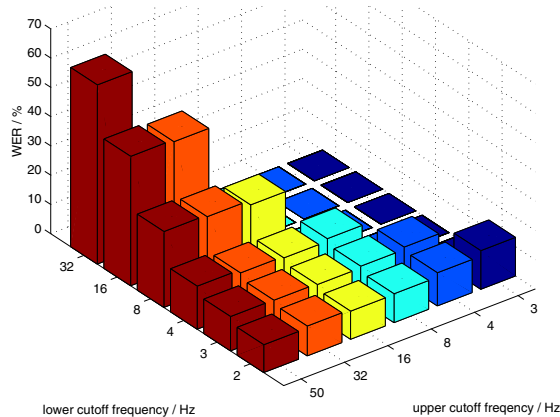


Figure 4: WER for isolated TIDigits in varying noise conditions measured for a FFNN classifier with 60 Gabor features. Each set was individually optimized on noisy training material. During optimization, the spectro-temporal features were restricted to a certain range of temporal modulation frequency between two cutoff frequencies denoted on the xy axes.

When further constraints are applied to the selection process, the ASR word error rate (WER) increases. In Fig. 4 the WER is given for automatically optimized sets with restricted ranges for temporal modulation best frequency  $\omega_t/2\pi$ . For TIDigits targets low temporal modulations between 2 and 4 Hz are most important, which matches the parameter statistics without constraints in Fig. 3.

## 6. Time will tell...

The new LSTFs extend existing feature types and are supported by recent research on auditory processing. First results are promising but a number of points remain open for further investigation. Most prominent is the search for an efficient automatic feature selection procedure that yields feature sets of satisfactory generality. In addition, the use of temporal modulation bandpass filters generally allows adjustment of the sampling rate according to the filter characteristics, which leads to the investigation of possible multi-rate features. Furthermore, it is very likely that LSTFs are even more suitable for syllable classification than for phoneme-based processing. This can be attributed to their syllable length and good performance on (mainly monosyllabic) digit corpora, and further promotes the use of syllables rather than phonemes as the basis for ASR.

## 7. References

- [1] H. Hermansky, "Should recognizers have ears?," *Speech Communication*, vol. 25, pp. 3–24, 1998.
- [2] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.
- [3] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. SAP*, vol. 2, no. 4, pp. 578–589, 1994.
- [4] H. Hermansky and S. Sharma, "TRAPS - Classifiers of temporal patterns," in *ICSLP*, 1998, vol. 3, pp. 1003–1006.
- [5] C.E. Schreiner and B.M. Calhoun, "Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions," *Auditory Neuroscience*, vol. 1, pp. 39–61, 1994.
- [6] 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.
- [7] L.M. Miller, M.A. Escabi, H.L. Read, and C.E. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory cortex," *J. Neurophysiol.*, vol. 87, pp. 516–527, 2002.
- [8] R. De-Valois and K. De-Valois, *Spatial Vision*, Oxford U.P., New York, 1990.
- [9] C. Kaernbach, "Early auditory feature coding," in *Contributions to psychological acoustics: Results of the 8th Oldenburg Symposium on Psychological Acoustics*. 2000, pp. 295–307, BIS, Universität Oldenburg.
- [10] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration," *JASA*, vol. 102, pp. 2906–2919, 1997.
- [11] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *JASA*, vol. 106, no. 5, pp. 2719–2732, 1999.
- [12] H. Fletcher, *Speech and Hearing in Communication*, Krieger, 1953, (There is a 1994 reprint ASA Edition.).
- [13] S. Greenberg, T. Arai, and R. Silipo, "Speech intelligibility derived from exceedingly sparse spectral information," in *ICSLP*, 1998.
- [14] S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," in *Eurospeech*, 2001.
- [15] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1, pp. 117–132, 1998.
- [16] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 2040–2050, 1999.
- [17] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, pp. 43–55, 1999.
- [18] C. Nadeu, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 1–2, pp. 93–114, 2001.
- [19] P. Somervuo, "Experiments with linear and nonlinear feature transformations in HMM-based phone recognition," in *ICASSP*, 2003.
- [20] P. Jain and H. Hermansky, "Beyond a single critical band in TRAP-based ASR," in *Eurospeech*, 2003, submitted.
- [21] H. Bourlard and N. Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," in *Adaptive Processing of Sequences and Data Structures*, vol. 1387 of *Lect. Notes in AI*, pp. 389–417. Giles, C.L. and Gori, M., 1998.
- [22] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*, 2000.
- [23] T. Gramß and H. W. Strube, "Recognition of isolated words based on psychoacoustics and neurobiology," *Speech Communication*, vol. 9, pp. 35–40, 1990.
- [24] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acustica united with acta acustica*, vol. 88, pp. 416–422, 2002.
- [25] K.P. Körding, P. König, and D.J. Klein, "Learning of sparse auditory receptive fields," in *Proc. IJCNN*, 2001.
- [26] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *ICSLP*, 2002.
- [27] T. Gramß, "Fast algorithms to find invariant features for a word recognizing neural net," in *IEEE 2nd International Conference on Artificial Neural Networks*, Bournemouth, 1991, pp. 180–184.