NON-STATIONARY FEATURE EXTRACTION FOR AUTOMATIC SPEECH RECOGNITION

Zoltán Tüske[1], *Pavel Golik*[1], *Ralf Schlüter*[1], *Friedhelm R. Drepper*[2]

[1]Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany
[2]Zentralinstitut für Elektronik, Forschungszentrum Jülich, 52425 Jülich, Germany
{tuske,golik,schluter}@cs.rwth-aachen.de, f.drepper@fz-juelich.de

ABSTRACT

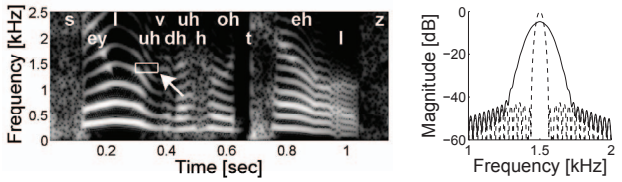
In current speech recognition systems mainly Short-Time Fourier Transform based features like MFCC are applied. Dropping the short-time stationarity assumption of the voiced speech, this paper introduces the non-stationary signal analysis into the ASR framework. We present new acoustic features extracted by a pitch-adaptive Gammatone filter bank. The noise robustness was proved on AURORA 2 and 4 tasks, where the proposed features outperform the standard MFCC. Furthermore, successful combination experiments via ROVER indicate the differences between the new features and MFCC.

Index Terms— non-stationary, pitch-adaptive, Gammatone, Gammachirp

1. INTRODUCTION

In state of the art automatic speech recognition (ASR) systems the most widely used acoustic features (Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficient (PLP)) are based on the Short-Time Fourier Transform (STFT). Although it is well know that speech contains non-stationary parts like stop consonants, the speech production model is described on a short-time scale ($\approx$ 30 ms) as a response of a linear time invariant (LTI) system to wide sense stationary or quasi-periodic excitation. In the case of voiced speech, the features are calculated from the harmonically structured power spectrum being composed of fundamental frequency (F_0) and its harmonics, where F_0 is assumed to be constant within the analysis window. In this case the Fourier analysis or comparable filter bank can be used to separate and extract the periodic modes containing the information of the vocal tract transfer function (VTTF). The time evolution of the fundamental frequency in particular challenges the stationarity assumption inside a short analysis window of about 30 ms. Figure 1(a) depicts a spectrogram showing the typical harmonic structure within voiced segments, exhibiting dynamic behavior due to F_0 variability.

The separation and reconstruction of non-stationary sinusoids is not possible by means of STFT. Figure 1(b) shows the magnitude spectrum of a stationary sinusoid compared to the one of a linear chirp. The dynamic of the latter results in a smearing of calculated energy among the bins. Further, it has been shown in psychoacoustic experiments, that the human auditory pathway benefits from adaptation to the time-varying frequency of harmonic modes: in presence of simultaneous speakers, the introduction of frequency modulation into voiced speech improves the intelligibility (*cocktail party effect*) [1].



(a) (b)

Fig. 1: (a) Spectrogram of "sale of the hotels" spoken by a woman, the chirp rate in white patch assumed in b. (b) Hamming windowed (25 ms) spectrum of chirped non-stationary harmonic mode (solid) compared with the stationary sinusoid spectra (dashed).

A non-stationary speech production model for voiced speech was introduced in [2]. The author also suggested an analysis suited to separate the harmonic modes and to extract their physical parameters. For this purpose adaptive Gammatone (GT) filters have been used.

In this paper we investigate the usage of non-stationary analysis [2] in the ASR framework replacing the standard windowed STFT in the MFCC feature extraction by pitch-adaptive GT filter bank. By suppressing the spectral region between the harmonics we are able to estimate the VTTF more robustly.

In Section 2 we summarize the related work. Section 3 gives an overview of the GT filters and their generalization for non-stationary harmonic signal analysis. The integration of the non-stationary filter bank into the MFCC pipeline is discussed in Section 4. Section 5 reports the experimental results on three different ASR tasks, while the conclusions are drawn in Section 6.

2. RELATED WORK

In [3] a model based on short-time stationary pitch-adaptive harmonics was used for noise robust VTTF estimation by suppressing the regions between the harmonics. The authors reported significant improvement on AURORA 2. To reduce the

interference between the signal periodicity and the window size, the authors in [4] substituted the classic fixed length windowing of the STFT by an F_0-adaptive window, which can also be interpreted as a variable bandwidth STFT. It slightly outperformed the conventional MFCCs. The properties of different approximations of GT filter transfer functions have been discussed in [5]. The stationary GT filter bank for acoustic feature extraction was investigated in [6]. The reported results are comparable to MFCC. However, in all these cases the authors assumed stationarity of the speech signal within each analysis frame.

3. FROM GAMMATONE TO GAMMACHIRP FILTER

In computational auditory models the peripheral filtering in the cochlea is typically described by GT filters as introduced in [7]. In the continuous time domain the impulse response is defined as:

$$f_\gamma(t) = A \cdot t^{\gamma-1} \cdot \exp(-2\pi \cdot f_b \cdot t) \cdot \cos(2\pi \cdot f_c \cdot t + \theta) \quad (1)$$

where γ denotes the filter order, f_b denotes a bandwidth parameter and f_c the center frequency in Hz.

A computationally efficient, complex valued, all-pole time domain implementation of a 4th-order linear GT filter in discrete domain was defined in [8] as a cascade of first-order filters. The difference equation of complex band-pass cascade element is:

$$\begin{aligned} y(n) &= x(n) + \alpha \cdot y(n-1) \\ \text{with } \alpha &= \lambda \cdot \exp\left(i \cdot 2\pi \cdot f_c / f_s\right) \end{aligned} \quad (2)$$

where $y(n)$ denotes the filter output and $x(n)$ the input at time n. The filter coefficient α is the complex pole of the filter where λ depends on the bandwidth parameter, f_c denotes the center frequency of the band-pass filter and f_s the sampling frequency.

In [2], the assumption of short-term periodicity of the voice source is dropped and the excitation is described instead as synchronized dynamics of time dependent part-tones. The author modelled the vocal tract excitation as time invariant non-linear filter response to a non-stationary input signal with varying frequency, referred to as the *fundamental oscillator*:

$$e(n) = Re\left\{ \sum_{k=1}^{K} c_k \cdot A(n) \cdot \exp(i \cdot k \cdot \phi(n)) \right\} \quad (3)$$

where $A(n)\exp(i \cdot \phi(n))$ corresponds to the fundamental oscillator and $\mathcal{F}(\cdot)$ produces the higher harmonics of $F_0 = \frac{f_s}{2\pi}\big(\phi(n) - \phi(n-1)\big)$.

Equation (3) introduces the higher harmonics of F_0. For the separation of harmonics with non-stationary frequencies the GT filter of Hohmann has been generalized and extended by the capability to adapt its center frequency $f_c(n)$ to the instantaneous chirp of the underlying modes:

$$\alpha(n) = \lambda \cdot \exp\left(i \cdot 2\pi \cdot f_c(n) / f_s\right) \quad (4)$$

The time dependency of the center frequency $f_c(n)$, being introduced into Equation (2), is inherited by the complex pole.

In [9] it has been shown that the band-pass feature of the GT filter is suited to isolate non-stationary harmonics. By choosing the center frequency of the GT filter identical to the instantaneous frequency of the isolated k-th harmonic, the phase of the filter output is identical to the phase of the input signal. The fundamental phase ϕ can be used to define the instantaneous fundamental frequency F_0 as well as the frequencies $f_c^{(k)}(n)$ of the band-pass filters

$$f_c^{(k)}(n) = k \cdot F_0(n) \quad \Rightarrow \quad \angle y_k(n) = k \cdot \phi(n) \quad (5)$$

The analysis is suited to extract harmonic modes with uncorrupted phases. In contrast to the reconstructed amplitude the reconstruction of the phase has no time delay. There is no restriction on the contour of the center frequency (except for continuity).

4. INTEGRATION INTO MFCC PIPELINE

To guarantee the continuity of the center frequency of a filter in the unvoiced regions (without valid F_0 estimation) we perform linear interpolation between voiced regions. Furthermore, covering every harmonic in the spectrum ($f_c < f_s/2$) leads to a varying number of filters with time dependent F_0. This issue is automatically solved by Mel triangular critical band integration which ensures constant feature space dimension for the further processing steps.

The windowed STFT used in many feature extraction methods can be expressed in terms of linear filtering operations as $STFT\{s(n), f\} = s(n) * \tilde{h}_f(n)$, where STFT of the signal $s(n)$ is calculated at time n and frequency f, and $\tilde{h}_f(n)$ corresponds to the STFT-equivalent filter: a time mirrored version of modulated window function (e.g. Hamming window). Usually the window size is about 30 ms, while the step width of about 10 ms corresponds to the downsampling of the filter output.

We replaced the windowing and STFT blocks in the MFCC extraction by a filter bank of stationary GT filters with the same parameters. In the second step we introduced the time dependency of the filter center frequencies according to the estimated pitch contour.

Further, after application of a non-stationary filter and an additional time average the output is downsampled. Because of the continuously changing filter center frequency the time averaging can not be performed independently from the spectral integration, therefore we used an approximation to separate the two steps and to reduce the computational costs:

$$\begin{aligned} M_d(n) &= \frac{1}{N+1} \sum_{m=n-\frac{N}{2}}^{n+\frac{N}{2}} \sum_{k=1}^{K} w_d(k \cdot F_0(m)) \cdot |y_k(m)| \\ &\approx \sum_{k=1}^{K} w_d\left(k \cdot \overline{F_0}(n)\right) \cdot \frac{1}{N+1} \sum_{m=n-\frac{N}{2}}^{n+\frac{N}{2}} |y_k(m)| \end{aligned}$$

where $M_d(n)$ denotes the output of the d-th component of a 20-dimensional Mel filter bank, $w_d(f)$ denotes the d-th

Table 1: Results on AURORA 2: Clean training

SNR	MFCC				NSGT			
[dB]	A	B	C	Avg.	A	B	C	Avg.
Clean	0.9	0.9	1.0	0.9	1.1	1.1	1.4	1.2
20	1.7	1.4	1.8	1.6	1.6	1.6	2.2	1.8
15	3.5	2.7	3.0	3.1	2.9	2.5	3.5	3.0
10	7.5	6.1	6.8	6.8	5.7	5.5	6.1	5.8
5	16.7	15.4	16.5	16.2	13.1	14.0	14.5	13.9
0	37.3	36.9	38.2	37.5	32.2	34.3	36.1	34.2
-5	69.1	69.5	68.1	68.9	66.5	66.9	64.5	66.0
Avg.	19.5	19.0	19.3	19.3	17.6	18.0	18.3	18.0
Rel. $\pm$					-9.7	-5.3	-5.2	-6.7

spectral weighting function (triangular filter), $y_k(n)$ corresponds to the filter output on the k-th harmonic of the estimated F_0 at time point n, i.e. $f_c^{(k)} = k \cdot \hat{F}_0(n)$. The average fundamental frequency over N samples is denoted by $\overline{F_0}(n) = \frac{1}{N+1} \sum_{m=n-\frac{N}{2}}^{n+\frac{N}{2}} \hat{F}_0(m)$.

Further processing in the MFCC pipeline is kept unchanged: the logarithm of the output of the triangular filter bank is decorrelated by application of the discrete cosine transformation (DCT) and normalized.

5. EXPERIMENTAL RESULTS

The new NSGT acoustic features were extensively evaluated and compared with MFCC on three different corpora: AURORA 2 and 4 and EPPS English task from TC-STAR 2007. Across all experiments, the RWTH ASR system was used as the recognizer system. The HMM topology is also the same in all setups, modelling each allophone by 3 states with repetitions, allowing loop, forward and skip transitions, while the silence consists of only one state. The acoustic model (AM) training is performed with respect to the maximum likelihood criterion. The emission probabilities were modelled via Gaussian mixtures with globally pooled diagonal covariance matrix.

5.1. AURORA 2

A set of experiments was conducted on the small vocabulary task AURORA 2. This corpus consists of US English digits under different types and levels of additive noise. The details on the corpus generation can be found in [10]. Test set A was used as development set, while the final evaluation was done by averaging all test sets over all 7 noise levels.

First, a baseline MFCC system was trained. The spoken digits are described by whole-word models, such that the number of HMM states is proportional to the number of phonemes per word. The AM was trained as follows: First, single Gaussians were estimated on MFCC+Δ+$\Delta\Delta$ features (45 dim.) using linear segmentation as initial alignment. After 7 density splits the final alignment was used for estimating a Linear Discriminant Analysis (LDA) matrix, mapping 9 consecutive time frames into a 45-dimensional space. With LDA transformed features the split-and-realign scheme was repeated, resulting in the final acoustic model.

Table 2: Results on AURORA 4: Detailed WERs of MFCC and NSGT systems, and their combination via ROVER

Mic.	System	Test set							Avg.
		1	2	3	4	5	6	7	
Sennh.	MFCC	3.8	8.7	11.9	18.1	17.6	15.2	20.3	13.7
	NSGT	4.9	9.7	13.2	17.8	16.6	14.7	18.9	13.7
	ROVER	3.6	7.4	11.5	15.5	14.6	12.8	16.5	11.7
Unk.	MFCC	16.3	20.9	35.2	37.4	39.3	33.1	37.3	31.4
	NSGT	14.8	23.5	30.8	33.7	34.3	30.3	33.6	28.7
	ROVER	13.7	19.8	29.0	30.5	32.8	28.4	31.6	26.5

In the NSGT system, a filter bank of linearly spaced GT filters with harmonics adapted center frequencies was employed. The search for an optimal constant bandwidth yielded the best recognition performance at 75 Hz, which corresponds to a coverage of approx. 45 % of the spectrum with 24 non-stationary filters. The results of these experiments for the systems trained on clean training data are given in Table 1. It can be easily seen, that the improvement originates mainly from the noisy parts, while on the clean test data, MFCC outperforms the NSGT features.

The second set of experiments on multi-conditional training data showed, that this advantage vanishes in case of a match between training and test data. The constant bandwidth had to be increased to 90 Hz to find an optimal value for new training data. Nevertheless, the average WER of conventional STFT-based features (13.5 %) could not be reached by the same NSGT approach (14.1 %).

5.2. AURORA 4

This corpus contains a closed vocabulary of 5000 words and is built from the WSJ0 recordings by adding real-world noise of different types and levels [11]. In contrast to AURORA 2, the words are modelled by phoneme sequences. The monophone training is performed in the same manner as described in the previous section. The resulting alignment is used for the estimation of an LDA matrix as well as a Classification And Regression Tree (CART) with 4000 leaves for triphone state-tying. In the recognition a 3-gram language model has been used.

The constant filter bandwidth in the NSGT system was chosen experimentally and set to 60 Hz, which corresponds to a coverage of less than the half of the effective bandwidth with average F_0 of 150 Hz. In both MFCC and NSGT systems, the output of 20 triangular Mel-scaled filters was decorrelated by a DCT, retaining only 16 components. The results of the experiments along with the STFT-based baseline MFCC system are shown in Table 2. The total WER averaged to 21.2 % (NSGT) and 22.5 % (MFCC). Again, the improvement originates mainly in the "hard" parts.

5.3. EPPS

The next experiments were conducted on EPPS English 2007. This LVCSR task contains 87.7 h training data, and the recognition is evaluated on the test sets *dev07* (3.1 h) and *eval07* (2.8 h). In contrast to both AURORA corpora, the performance of a similar NSGT system could not reach the base-

Table 3: Word error rates on EPPS English 2007

System	MFCC	NSGT				ROVER
		$f_c^{(k)}$:	$k \cdot F_0$	$k \cdot F_0$	$k \cdot F_0/2$	
		$f_b^{(k)}$:	75 Hz	120 Hz	$F_0/2$	
dev07	17.3		19.6	19.5	18.5	16.9
eval07	16.2		19.2	18.6	17.1	15.8

line MFCC setup. Increasing f_b up to 120 Hz had only small impact on the performance. Hence, the coverage of the spectrum was improved by putting additional GT filters between the integer multiples of the estimated fundamental frequency. The bandwidth of the filters was adapted to one half of the estimated $F_0(n)$, resulting in an almost complete filter bank. This filter bank design led to more competitive results, which are summarized in Table 3. In the ROVER [12] experiments, MFCC was combined with the best NSGT system.

5.4. Discussion

The observed noise robustness on both AURORA corpora can be explained by the reduced coverage of the spectrum. Following harmonic contours by narrow band-pass filters can also be considered as suppression of the regions between the harmonics. These regions have been shown to be more vulnerable to noise than the harmonic peaks [13]. The NSGT features thus reduce the mismatch between the clean training and the noisy test data by a more robust estimation of the spectral envelope than STFT-based MFCC.

In addition to the baseline MFCC systems, a set of experiments was performed by using stationary GT filter banks without F_0 adaptation. This allows to compare the adaptive and non-adaptive filtering under the same conditions and hence to isolate the influence of dynamic filter parameters. The filter bank was designed identical to the STFT as described in Section 4. The resulting averaged error rates of 19.2 % (clean) and 13.0 % (multi-conditional training) on AURORA 2 were even slightly better than standard MFCC. On AURORA 4 the performance of STFT-based features could also be repeated with 22.2 % WER. Reducing the filter bandwidth or spacing without adaptation to F_0 led to a degradation as expected.

Although the amplitude-based NSGT features were derived from the MFCC pipeline, the combination of both recognizer outputs via ROVER shows further improvement in all experiments, indicating differences in the acoustic features.

6. CONCLUSIONS

In this work, a novel approach of non-stationary signal analysis based on pitch-adaptive GT filters was presented. The extracted model parameters could be integrated into acoustic feature extraction for ASR. In the experiments on AURORA 2 and 4, the amplitude-based features exhibited noise robustness and yielded competitive results with the state of the art MFCC systems. The output phase of the complex GT filters could be interpreted in terms of an appropriate underlying physical model, such that a connection to phonetic events could be established systematically.

In our future work we will concentrate on a better handling of unvoiced segments, e.g. by using an additional stationary voicedness-based fallback filter bank.

7. ACKNOWLEDGMENT

This work has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement No. 213850.11, SCALE.

8. REFERENCES

[1] S. McAdams, "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," *Journal of the Acoustical Society of America*, vol. 86, no. 6, pp. 2148–59, Dec. 1989.

[2] F. R. Drepper, "A two-level drive-response model of non-stationary speech signals," *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, April 2005.

[3] M. L. Seltzer, J. Droppo, and A. Acero, "A harmonic-model-based front end for robust speech recognition," in *EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 1277–1280.

[4] G. Garau and S. Renals, "Pitch adaptive features for LVCSR," in *INTERSPEECH*, Brisbane, Australia, Sept. 2008, pp. 2402–2405.

[5] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, "Introducing the differentiated all-pole and one-zero Gammatone filter responses and their analog VLSI log-domain implementation," in *Proc IEEE Int. Midwest Symposium on Circuits and Systems*, San Juan, Puerto Rico, Aug. 2006, pp. 561–565.

[6] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *ICASSP*, Honolulu, HI, USA, April 2007, pp. 649–652.

[7] E. de Boer, "Synthetic whole-nerve action potentials for the cat," *Journal of the Acoustical Society of America*, vol. 58, no. 5, pp. 1030–1045, Nov. 1975.

[8] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, May 2002.

[9] F. R. Drepper and R. Schlüter, "Non-stationary acoustic objects as atoms of voiced speech," in *34. Jahrestagung für Akustik der Deutschen Gesellschaft für Akustik*, Dresden, Germany, March 2008, pp. 249–250, [corrected version: Eq. (5)].

[10] D. Pearce and H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA Tutorial and Research Workshop ASR2000*, Paris, France, Sept. 2000, pp. 181–188.

[11] N. Parihar and J. Picone, "*Aurora Working Group: DSR Front End LVCSR Evaluation*," Technical Report, Mississippi State University, MS, USA, Dec. 2002.

[12] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, California, USA, Dec. 1997, pp. 347–354.

[13] Q. Zhu and A. Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," *Computer Speech and Language*, vol. 17, no. 4, pp. 381–402, Oct. 2003.