Non-Stationary Signal Processing and its Application in Speech Recognition

Zoltán Tüske, Friedhelm R. Drepper, Ralf Schlüter

Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany

{tuske,drepper,schluter}@cs.rwth-aachen.de

Abstract

The most widely used acoustic feature extraction methods of current automatic speech recognition (ASR) systems are based on the assumption of stationarity. In this paper we extensively evaluate a recently introduced filter stable, non-stationary signal processing method, which relies on an adaptive part-tone decomposition of voiced speech to obtain alternative feature vectors for ASR. The non-stationary filterbank allows for more noise robust amplitude based features by suppressing the between-harmonics regions. Furthermore, by adapting the center filter frequencies to the underlying acoustic modes, it is possible to obtain useful phase features which can be interpreted in terms of the non-stationary dynamics within the vocal tract. The features are evaluated on different tasks ranging from vowel classification up to large vocabulary continuous speech recognition.

Index Terms: non-stationary, adaptive filter, noise robust, phase features, ASR

1. Introduction

Assuming weak stationarity, the time-frequency decomposition of speech signals is mainly based on Short-Time Fourier Transform (STFT). Although it is well known, that speech contains non-stationary parts, the speech production model is described on a short-time scale (about $30ms$) as response of a linear time invariant system to wide sense stationary or quasi-periodic excitation.

However, the relevant components of the phonetic inventory of human languages are characterized also by various non-stationary processes (e.g. word accents, diphthongs), where the underlying physiological processes generate characteristic amplitude- and frequency modulation. Tonal languages increase this list by different phonetic interpretation of pitch contours. Thus, the time evolution of the fundamental frequency, in particular, challenges the stationarity assumption even inside a short analysis window.

Analyzing chirped sinusoids with STFT leads to typical smearing effects in the amplitude spectrum [1]. Furthermore, as was shown in psychoacoustic experiments, the introduction of frequency modulation into voiced speech is beneficial to the intelligibility in the presence of simultaneous speakers, indicating that the human auditory system can profit from adaptation to the time-varying harmonic modes [2].

The relevance of pitch for intra-species communication, in particular for voiced speech of humans, suggests that it might be advantageous for the analysis of voiced speech to replace the a priori choice of filter frequencies by a closed loop adaptation to the input signal.

In [3] a STFT comparable filterbank was generalized by introducing time-dependent filter frequencies and an iterative adaptation of the filter frequency contours. The adaptation leads to non-stationary bandpass filters suited to generate filter outputs with uncorrupted phases. The filter frequency contours were obtained as stable fixed point (asymptote) of the iterative update of the filter frequency contours by using frequency contours of the filter outputs. When choosing appropriate bandwidths such filters can e.g. be used to extract uncorrupted phases of the underlying harmonic modes of a vowel.

The present paper replaces the STFT based amplitudes of current ASR systems by amplitude outputs of appropriately adapted non-stationary gammatone filters. As a more innovative step, the amplitude based part of the acoustic feature vector is supplemented by phase differences of neighbouring part-tones. We demonstrate that such phase cues are suited to detect the phase jumps which result from the passage of a harmonic mode through the resonance of a formant. The novel feature vector is extensively evaluated in vowel classification, phoneme recognition, and small and large vocabulary speech recognition tasks. Moreover, the noise robustness is tested on standard tasks for noisy speech.

The paper is organized as follows, Section 2 gives a short overview of non-stationary bandpass filters. The filter implementation for non-stationary signal analysis is discussed in Section 3. The details of non-stationary features are presented in Section 4. The extensive experimentation with the derived features are carried out in Section 5. The paper closes with conclusions.

2. Non-Stationary Bandpass Filter

The time-frequency decomposition of signals by bandpass filters with a priori fixed center frequencies can be generalized by introducing time dependent center filter frequencies (CFF) $\omega_j(t)$. Denoting the input signal as $S(t)$, the envelope of the impulse response of the jth bandpass filter as $W_j(t)$ and the time delay of the envelope maximum of the impulse response as τ_j, the response of jth complex subband $X_j(t)$ can be expressed as

$$X_j(t) = \int_{-\infty}^{t+\tau_j} S(\tau)W_j(t-\tau)\, exp\left(i\int_{\tau}^{t}\omega_j(\tau')d\tau'\right)d\tau. \tag{1}$$

The instantaneous phase of $X_j(t)$ will be denoted as $\varphi_j(t)$. As was shown in [4], it is possible to obtain the uncorrupted instantaneous phase $\Phi_j(t)$ of a non-stationary sinusoid (which may be interpreted e.g. as a well separable single (jth) harmonic mode of a vowel), if one succeeds to adapt the CFF to the instantaneous frequency of the underlying mode:

$$\varphi_j(t) = \Phi_j(t) \quad \text{if} \quad \omega_j(t) = \dot{\Phi}_j(t), \tag{2}$$

where $\Omega_j(t) \doteq \dot{\Phi}_j(t)$ will denote the instantaneous frequency of the analyzed sinusoid. Furthermore, it was shown that a mapping function F of the filter frequency contour can be defined which leads to the desired equality of Eq. (2). As indicated in Eq. (1) this mapping transforms a filter frequency contour $\omega_j(t)$ within the analysis window $t' \leq t \leq t''$ to the instantaneous filter output phase contour $\varphi_j(t)$. Together with the time derivative operator d/dt the mapping obtains an identical input and output range and can thus be iterated

$$\omega_j^{(n+1)}(t) := \frac{d\varphi_j^{(n)}}{dt} = \frac{d}{dt}F\left\{\omega_j^{(n)}(t)\right\} \quad \text{for} \quad t' \leq t \leq t''. \tag{3}$$

The left hand side of (3) denotes the updated CFF for the $n+1$. iteration step obtained from output frequency ($d\varphi_j^{(n)}/dt$) of the nth iteration step, whereas the right hand side expresses the filtering with the CFF contour $\omega_j^{(n)}(t)$, resulting in filter output $X_j^{(n)}(t)$. Eq. (3) has a fixed point — an invariant CFF contour — given as $\omega_j^{(\infty)}(t) = \Omega_j(t)$. As a characteristic feature of this fixed point, three frequency contours are identical: the one of the CFF, $\omega_j^{(\infty)}(t)$, the one of the output frequency, $d\varphi_j^{(\infty)}(t)/dt$, and the one of the input frequency, $\Omega_j(t)$. The invariant filter frequency contour generates an output phase which is identical to the phase of the input: $\varphi_j^{(\infty)}(t) = \Phi_j(t)$.

For gaussian type impulse responses $W_j(t)$ and differentiable input frequency contours which can be well approximated by a linear function (linear chirp), it could be shown that the convergence in the neighbourhood of the fixed point is a stable one and converges with the *third* power of the deviation from the fixed point [5]. Such a fixed point is often called as super stable fixed point.

All input signals with a phase velocity $\Omega_j(t)$ which is substantially different from a given CFF contour $\omega_j(t)$ experience a damping due to interference. Due to this bandpass property, it is expected that a broadband input signal, given e.g. as a superposition of several sinusoids with sufficiently separated or different frequency contours, will result in several stable fixed points, each with a finite size basin of attraction.

3. Gammachirp Filter

The well known Gammatone (GT) bandpass filters are ideally suited to be generalized towards non-stationary signal processing. A computationally efficient, complex valued, all-pole approximation of a GT filter in the discrete time domain was defined in [6] as a cascade of first-order filters. The difference equation of complex bandpass cascade element is

$$\begin{aligned} X_j[k] &= S[k] + \alpha_j \cdot X_j[k-1] \\ \text{with} \quad \alpha_j &= \lambda \cdot \exp\left(i\,\omega_j T\right), \end{aligned} \tag{4}$$

where $X_j[k]$ denotes the jth filter output, and $S[k]$ the sampled input signal. The filter coefficient α_j is the complex pole of the first order filter, where λ depends on the bandwidth parameter, ω_j denotes the center frequency of the jth bandpass filter, and $T = 1/f_s$ the sampling period. For the separation of harmonics with non-stationary frequencies the GT filter of Hohmann was generalized in [3] and extended by the capability to adapt its center frequency ω_j to the instantaneous chirp of the underlying mode

$$\alpha_j[k] = \lambda \cdot \exp\left(i\,\omega_j[k]T\right). \tag{5}$$

The time dependency of the center frequency $\omega_j[k]$, being introduced into Eq. (4), can be interpreted as a computationally efficient time discrete implementation of gammachirp filter of Eq. (1). Expressing the filter phase of the bandpass filter in Eq. (1) in discrete time domain results in

$$\begin{aligned} \left.\int_{\tau}^{t} \omega_j(\tau')d\tau'\right|_{\substack{t=kT \\ \tau=k'T}} &= \sum_{k''=k'+1}^{k} \int_{(k''-1)T}^{k''T} \omega_j(\tau')d\tau' \\ &= T \sum_{k''=k'+1}^{k} \omega_j[k''], \end{aligned} \tag{6}$$

where $\omega_j[k]$ is introduced as the average filter frequency between two samplings. Assuming piece-wise linear filter frequency, $\omega_j[k]$ can be expressed as:

$$\omega_j[k] = \omega_j\left(T\left(k - \frac{1}{2}\right)\right) \tag{7}$$

$$\text{if} \quad \omega_j(t) = \omega_j(t') + \dot{\omega}_j \cdot (t - t') \quad \text{and} \quad t' < \{t, kT\} < t''.$$

Thus, the momentary frequency of the discrete non-stationary Gammatone filter reflects the average frequency between two samples instead of the sampled continuous filter frequency. In case of linear gammachirp filter this corresponds to a half sample correction term in the sampled version of $\omega_j(t)$.

4. Features Derived from Non-Stationary Signal Processing

For voiced speech signals as input and sufficiently narrow bandwidths of the filters ($\sim 60Hz$) the basins of attraction of the filter stabilization process can be expected to have a harmonic structure in the sense that different filter-stable frequency contours can be reached from different harmonic start contours. Furthermore, since the convergence of the filter stabilization process is extremely fast, we can expect that a single iteration step might be sufficient. To obtain the harmonic start contours $\omega_j^{(0)}(t) = j \cdot F_0(t)$, a conventional fundamental frequency estimator [7] is applied to extract $F_0(t)$. For the iterative filter update process the output frequencies of the filters are approximated by linear chirp in $40ms$ analysis window, and the new parameters are estimated according to [8]. For the unvoiced regions without valid F_0 estimation, linear interpolation between voiced regions is performed. Furthermore, the filter update is not executed in unvoiced regions.

4.1. Amplitude features

In order to integrate the amplitude output of non-stationary filters into the standard MFCC feature extraction, the windowing

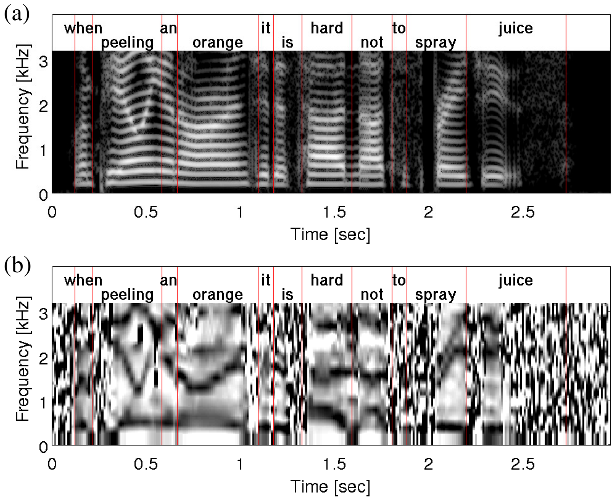

Figure 1: (a) Spectrogram and (b) phasegram derived from the reconstructed part-tones of an utterance from TIMIT database

and STFT block is replaced by a bank of GT filters with time dependent center filter frequencies [1]. Covering every harmonic in the spectrum, $\omega_j(t) < \omega_{\text{nyquist}}$ leads to a varying number of filters. This issue is automatically solved by the Mel triangular critical band integration which ensures constant feature space dimension. According to the MFCC pipeline, the further processing steps correspond to the application of logarithm, discrete cosine transformation (DCT), and segment-wise mean and variance normalization.

4.2. Phase features

As shown in [9], the time evolution of appropriately chosen relative phase shifts of harmonics shows a characteristic pattern which can often be interpreted in terms of underlying formant resonances of vowels. The phase features of the present study are chosen as (wrapped) differences of the relative phases of neighboring filter-stable part-tones:

$$\begin{aligned}\Delta\varphi_j(t) &= \varphi_j(t) - j\varphi_1(t) - (\varphi_{j-1}(t) - (j-1)\,\varphi_1(t)) \\ &= \varphi_j(t) - \varphi_{j-1}(t) - \varphi_1(t). \qquad (8)\end{aligned}$$

To obtain constant feature dimension at every time point, $\Delta\varphi_j(t)$ is interpolated onto a fixed set of frequencies having the same resolution as the STFT.

Fig. 1 shows the Hamming windowed STFT amplitude spectrum and the corresponding phasegram of an utterance from the TIMIT corpus. In the regions where a well defined F_0 is available, the typical formant structure of the utterance can easily be recognized from the phasegram. Based on the observation that stable patterns of the phasegram are mainly limited to the lower frequency range, the phase feature extraction is limited to the range up to $2500Hz$. Representing a complementary cue, the relative phase of harmonic neighbors can be expected to provide a promising supplementary acoustic feature for speech recognition.

5. Experiments

5.1. Vowel classification and phoneme recognition

5.1.1. Experimental setup

As a first investigation of phase and amplitude features extracted from filter-stable part-tones, vowel classification and phoneme recognition experiments are performed on the TIMIT corpus. Like in many previous studies the original 61 phones of TIMIT [10] are mapped to a smaller subset of 39 phones.

For vowel classification only those feature vectors are considered, which are labelled as one of the 14 vowels in the manual alignment. Keeping the number of parameters constant, the Multi Layer Perceptron (MLP) based vowel classifier is trained on nine consecutive frames of the features. The 3-layer MLPs are trained using cross-entropy criterion and approximate phoneme class posterior probabilities. As training set the standard 3696 'si' and 'sx' sentences are chosen resulting in ∼273000 feature vectors. Ten percent of the training set (chosen randomly) is used as cross validation (CV) set for adjusting the learning rate and to prevent overfitting. Whereas the Core test contained ∼22000, the Full test set consisted of ∼160000 vowel related feature vectors. The classification results of the present study are obtained by classifying every feature vector of each vowel token independently. For comparison with previous work (e.g [11]), the classification errors achieved at the midpoint of each vowel token are also reported.

To perform the phoneme recognition on the TIMIT database, Gaussian Mixture based acoustic Models are trained on concatenated MFCC and TANDEM [12] features, where the single hidden layer MLP based posterior features are trained using either phone or phone state target. During the recognition a bigram phoneme language model estimated on the training data is used.

5.1.2. The effect of bandwidth and noise level

In the first experiments the filterbank is not iterated, however, the F_0 contour used for the CFF of the higher harmonics is extracted on clean data. To analyze the interaction between the filter bandwidth and F_0, the training and test corpora are split into male (lower F_0) and female (higher F_0) parts and are used to train separate classifiers by filtering the data with different bandwidth. About 30% of all vowel related feature vectors belong to female speakers. Furthermore, different levels of white noise are added to both the training and test recordings to investigate the noise sensitivity of the extracted features. The maximum quantization level of the input signal corresponds to $0dB$.

As can be seen in Table 1, the MFCC features clearly outperform the non-stationary filter based phase features (rel.NSGT phase). However, the classification error rates achieved by phase features alone without any energy related features are remarkable. Since the phase features are not post processed in this experiment, their results are also compared to the STFT amplitude spectrum (log(|STFT(.)|)). The experiments reveal that the usefulness of the phase features defined in Eq. 8 decreases with increasing noise level, especially in case of male speakers (low F_0). The experiments also show that narrow band filtering is more beneficial in noise and that the bandwidth becomes less important for higher F_0. In case of male speakers (lower F_0) narrow band filtering should be applied to avoid interference form the neighboring harmonics, which could degrade the classification performance.

Table 1: *Vowel classification error [%] achieved with relative phase difference (8) of the output of the non-stationary filters, results are compared to standard features*

Added noise level [dB]	Gender	Features						
		rel.NSGT phase					log(|STFT(.)|)	MFCC
		Bandwidth [Hz]						
		40	60	80	100	120		
None	F	50.0	49.5	49.0	48.5	49.3	37.7	34.0
	M	44.8	44.0	45.9	50.6	56.5	30.7	28.7
-60	F	52.0	52.0	52.0	54.1	55.1	39.3	32.5
	M	50.8	54.4	58.6	60.2	62.4	31.5	28.8
-40	F	71.6	71.5	73.7	74.1	75.2	54.1	45.8
	M	71.6	72.2	72.7	73.4	73.8	44.0	40.0

Table 2: *Vowel classification error [%] achieved with the logarithmized amplitude output of the non-stationary filter, results are compared to standard features*

Added noise level [dB]	Gender	Features						
		log(|NSGT(.)|)					log(|STFT(.)|)	MFCC
		Bandwidth [Hz]						
		40	60	80	100	120		
None	F	41.0	40.2	39.7	39.4	38.8	37.7	34.0
	M	33.9	33.3	32.2	32.6	32.8	30.7	28.7
-60	F	40.2	39.3	39.7	39.2	39.7	39.3	32.5
	M	34.0	32.9	33.7	33.2	33.3	31.5	28.8
-40	F	49.5	49.0	49.3	49.6	49.9	54.1	45.8
	M	43.2	42.9	43.2	44.5	45.9	44.0	40.0

Table 3: *Vowel classification error [%] on TIMIT database using non-stationary phase and standard amplitude based features (results achieved by classifying exclusively at the midpoint of each vowel are indicated in brackets)*

Test	rel.phase		log(|STFT(.)|)	MFCC	MFCC +DCT(rel.phase)
	filter update				
	no	yes			
CV	48.9 (44.3)	49.4 (45.3)	35.5 (31.3)	33.2 (29.4)	31.2 (27.6)
Core	50.8 (46.7)	50.4 (46.1)	38.4 (34.1)	35.5 (31.1)	34.1 (29.8)
Full	49.4 (45.2)	49.4 (45.2)	37.2 (33.0)	34.7 (30.4)	33.1 (29.2)

Table 4: *Comparison of GMM based phoneme recognition error rates using only MFCC or combined MFCC+phase based MLP posterior features, where MLP^{p} and MLP^{ps} indicate phone or phone state posteriors, respectively*

AM	Features	Test	
		Core	Full
monophone	MFCC	31.1	29.7
	MFCC+MLP^{p}(MFCC)	29.7	28.2
	MFCC+MLP^{p}(MFCC+phase)	28.6	27.5
triphone (850)	MFCC	28.6	28.2
	MFCC+MLP^{p}(MFCC)	28.6	28.1
	MFCC+MLP^{p}(MFCC+phase)	27.4	27.0
	MFCC+MLP^{ps}(MFCC)	27.8	27.0
	MFCC+MLP^{ps}(MFCC+phase)	27.2	27.0

The experiments were also repeated with the amplitude output of the non-stationary gammatone filterbank (log(|NSGT(.)|)) The results can be seen in Table 2. Similarly to the phase features the logarithmized output of the filterbank is interpolated onto fixed frequencies. The amplitude output of the non-stationary gammatone filterbank achieves comparable results to logarithmized STFT amplitude. Furthermore, the log(|NSGT(.)|) features depend less sensitively on the bandwidth than the phase. In case of higher noise (-40dB) level the suppression of between harmonics regions is clearly advantageous, and the log(|NSGT(.)|) outperforms log(|STFT(.)|), indicating the noise robust property of narrow band F_0 synchronous signal processing. However, for the clean condition the suppression of between-harmonics regions leads to filtering out useful information, and thus to performance degradation. Using constant bandwidth, the higher F_0 leads to wider between harmonics regions, thus the female results show more bandwidth sensitivity in clean speech condition.

5.1.3. Vowel classification

Considering the results in Section 5.1.2, a filter bandwidth of 60Hz is selected for the further experiment. The effect of the filter stabilization is investigated by using the phase features. The results based on the complete training set (without splitting up) are shown in Table 3. As can be seen, the classical mean and variance normalized cepstral coefficients (MFCC) achieve the best accuracy. As a more fair comparison between phase and amplitude based features, the phase features are also compared with the logarithmized amplitude of short-time Fourier spectrum (log(|STFT(.)|)). Again, there is no doubt that the phase based cues alone perform worse than MFCC or the amplitude features, however the phase based results deserve attention. Furthermore, there is insignificant difference between the results of phase features with or without additional filter update. The fact that the phase features do not show improvement after filter update underlines that a frequency contour of the filter output is much more accurate than the corresponding center filter frequency contour. Since the iterative filter update does not lead to performance improvement, in the following the integer harmonics of the F_0 estimation are used as center filter frequency contours.

In addition, Table 3 shows results achieved with the concatenated MFCC and phase features, as well. For this experiment the dimension of the phase features is reduced to 8 by DCT (optimized on CV set). Comparing to MFCC alone, the concatenated features lead to a 4% relative decrease in the classification error. The difference between the two classifiers is statistically significant at 99.5% confidence level on the full test using McNemar's test.

5.1.4. Phoneme recognition

The phase features are further tested on a phoneme recognition task. Using Gaussian Mixture Model (GMM) based Hidden Markov acoustic Models (HMM), the phase features are integrated with the TANDEM [12] approach into the feature extraction. As can be seen in Table 4, the non-stationary gammatone phase features improve the recognition performance if monophone acoustic models are used, however, if the more complex tied-triphone state based modeling with phone state label based MLP features are applied, the phase features cannot contribute to phoneme recognition error reduction on the full test set.

5.2. Noisy ASR experiments

Based on the observation of Section 5.1.2, the noise robustness of amplitude features extracted from non-stationary bandpass filterbank is investigated by conducting noisy speech recognition experiments on Aurora 2 and Aurora 4 tasks. The amplitude features are extracted according to Section 4.1 and are denoted in the followings as NSGT. Since the performance of the relative phase of the non-stationary filterbank is highly sensitive to the noise level, the phase features are discarded in these experiments.

5.2.1. Experimental setup

The GMM based acoustic models are trained using maximum likelihood criteria. Instead of the recognizer defined in [13], another publicly available recognizer is used [14]. In the digits string (Aurora 2) recognition experiments, the digits are described by whole-word models, such that the number of HMM states is proportional to the number of phonemes per word. Whereas in the intermediate size vocabulary continuous speech recognition task, Aurora 4, the words are modelled by 4500 tied triphone states, the recordings sampled at $16kHz$ are used, and a trigram language model is applied during the recognition.

5.2.2. Results

The recognition performance of the different features are given in Table 5. Instead of reporting the average results achieved over signal-to-noise ratio range between 0 and $20dB$, the total average is shown. As can be seen, the NSGT features outperform the MFCC on the noisy parts, while — as it is observed on TIMIT in Section 5.1.2 — the suppression of the regions between the harmonics deteriorates the results in case of clean data. Optimization of the bandwidth parameter on test set A resulted in $75Hz$ bandwidth, which corresponds to a coverage of approx. 45% of the spectrum. Experiments on multi-conditional training data show, that the advantage of NSGT over MFCC vanishes, and that the MFCC slightly outperformed the NSGT, suggesting that the generalization of the acoustic model cannot be improved by our noise-suppression method.

In continuous intermediate size vocabulary experiments similar observations can be made, the results are shown in Table 6. The improvement originates mainly from the noisy part of the corpus (Test 3 - Test 7). In order to isolate the influence of the dynamic filter parameter, a further set of experiments is performed on Aurora 4. The STFT filterbank in the MFCC pipeline was substituted by a set of stationary GT filters (MFCC$_{\text{GT}}$). As can be seen, the MFCC$_{\text{GT}}$ features perform similar to MFCC, therefore the improvement achieved by NSGT features clearly relates to the F_0 synchronous non-stationary signal processing. Moreover, ROVER [15] based system combination indicates complementarity between the MFCC and NSGT features leading to improvements on all test sets.

Table 5: *Word error rates [WER] achieved on Aurora 2 using clean training data and different features*

SNR	MFCC			NSGT		
[dB]	A	B	C	A	B	C
Clean	0.9	0.9	1.0	1.1	1.1	1.4
20	1.7	1.4	1.8	1.6	1.6	2.2
15	3.5	2.7	3.0	2.9	2.5	3.5
10	7.5	6.1	6.8	5.7	5.5	6.1
5	16.7	15.4	16.5	13.1	14.0	14.5
0	37.3	36.9	38.2	32.2	34.3	36.1
-5	69.1	69.5	68.1	66.5	66.9	64.5
Avg.	19.5	19.0	19.3	17.6 (-9.7)	18.0 (-5.3)	18.3 (-5.2)

Table 6: *Detailed WER results on AURORA 4 using different features. MFCC$_{GT}$ indicate the substitution of STFT filterbank with stationary Gammatone filters in the MFCC pipeline, whereas ROVER combines MFCC and NSGT systems*

		System	Test set							Avg.
			1	2	3	4	5	6	7	
Microphone	Sennheiser	MFCC	3.8	8.7	11.9	18.1	17.6	15.2	20.3	13.7
		MFCC$_{\text{GT}}$	4.0	9.0	13.0	18.5	18.0	14.7	21.2	14.1
		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
	Unknown	MFCC	16.3	20.9	35.2	37.4	39.3	33.1	37.3	31.4
		MFCC$_{\text{GT}}$	16.1	20.2	35.3	35.9	38.2	33.7	38.2	31.1
		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

5.3. Large vocabulary speech recognition

In the final experiments we investigate the non-stationary features in an English large vocabulary continuous speech recognition task. The collected data of the European Parliament Plenary Sessions (EPPS) are recorded in clean condition with professional equipment, therefore the phase features are also tested.

5.3.1. Experimental setup

To train the GMM acoustic model containing 4500 generalized triphone states 88 hours of speech data are used. The acoustic models consist of Gaussian mixture distributions with a globally pooled diagonal covariance matrix. The performance of the final systems are evaluated on the development (Dev07) and evaluation (Eval07) data of 2007. Each corpus contains 3 hours of audio data, and the development corpus is used for tuning. Instead of training the mixtures from scratch, an alignment created by a previous system is applied to initialize the 6-state left-to-right HMMs.

5.3.2. Results

To investigate the complementarity of the NSGT features to the standard stationary signal-processing based features, the following features are extracted: MFCC, PLP [16], GT [17]. As the results in Table 7 show, the NSGT system performs slightly worse than the MFCC system. Considering the fact that this recognition task is based on clean data, the results confirm again that the suppression of between harmonics region is undesired in clean condition. Nevertheless, ROVER based system com-

Table 7: *Comparison and ROVER combination results of standard and NSGT features on EPPS 2007 corpora*

Features				Test	
MFCC	GT	PLP	NSGT	Dev07	Eval07
X				17.3	16.2
	X			17.6	16.8
		X		17.8	16.8
			X	18.3	17.5
X	X			16.3	15.3
X		X		16.5	15.6
X			X	16.8	15.9
X	X	X		16.2	15.1
X	X	X	X	16.0	15.0

Table 8: *Results on EPPS 2007 corpora, using supplementary non-stationary signal processing based phase features in TANDEM approach*

Features	Test	
	Dev07	Eval07
MFCC	17.3	16.2
MFCC+MLP(MFCC)	16.2	14.9
MFCC+MLP(MFCC+phase)	16.5	15.0

binations show that NSGT and MFCC can lead to performance improvement. As can also be seen, using NSGT as fourth system improves the combination results of the three standard features further, although the gain is not high.

Table 8 shows experiments using phase features, where the relative phase shifts of neighboring part-tones are integrated into the feature extraction according to Sec. 5.1.4. The experiments are carried out with 3-layer MLP based posterior features trained on phoneme target labels only. The phase features do neither improve nor deteriorate the final recognition performance measured on the Eval07 set.

6. Conclusions and Future Directions

In this study a recently proposed non-stationary signal processing technique was evaluated on numerous speech recognition tasks. Investigating the interaction between F_0 range, noise level, and bandwidth, amplitude and phase based features were designed to extract physical parameters of the acoustic dynamics of voiced speech. Using a bank of narrow band non-stationary filters, the experiments revealed that a precise conventional estimation of the F_0 contour and its higher harmonics is sufficient to obtain the non-stationary filter frequency contours, thus making the filter update unnecessary. Having tested the derived features in many speech recognition applications, we can conclude, that the amplitude output of the non-stationary filterbank can result in noise robust features for situations with mismatched training and testing data. Considering the phase features as supplement to amplitude based cues, improvement observed in low-level classification experiments did not generalize to large vocabulary continuous speech recognition.

As future direction, since the separation of harmonics with non-stationary filters was successfully demonstrated, there is also a hint that the time variant filters might be useful to support speech separation. Inspired by the success of the STFT, the present study is based on filterbanks with STFT like non-audiological bandwidths. The investigation of the convergence behavior of filters with audiological bandwidth covering several sinusoids could be also part of the further research.

Acknowledgement

This work has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. [213850]. 11, Speech Communication with Adaptive Learning - SCALE.

7. References

[1] Z. Tüske *et al.*, "Non-stationary feature extraction for automatic speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 5204–5207.

[2] S. McAdams, "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," *J. Acoust. Soc. Am.*, vol. 86, no. 6, pp. 2148–59, Dec. 1989.

[3] F. R. Drepper, "Voiced speech as response of a self-consistent fundamental drive," *Speech Communicaton*, vol. 49, no. 3, pp. 186–200, 2007.

[4] F. R. Drepper and R. Schlüter, "Non-stationary acoustic objects as atoms of voiced speech," in *Proc. of DAGA*, 2008, pp. 249–250.

[5] Z. Tüske *et al.*, "Phase difference of filter-stable part-tones as acoustic feature," in *Proc. of IEEE Statistical Signal Processing Workshop*, 2012, p. accepted for publication.

[6] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica / Acustica*, vol. 88, pp. 433–442, May 2002.

[7] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. of Int. Conf. on Spoken Language Processing*, 2000, pp. 464–467.

[8] S. Kay, "A fast and accurate single frequency estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 1987–1990, 1989.

[9] I. Saratxaga *et al.*, "Using harmonic phase information to improve ASR rate," in *Proc. of Interspeech*, 2010, pp. 1185–1188.

[10] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[11] S. A. Zahorian and Z. B. Nossair, "A partitioned neural network approach for vowel classification using smoothed time/frequency features," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 4, pp. 414 –425, 1999.

[12] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 2000, pp. 1635–1638.

[13] H. Hans-Günter and P. David, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ASR-2000*, Paris, France, Sept. 2000, pp. 181–188.

[14] D. Rybach *et al.*, "RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[15] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 1997, pp. 347–354.

[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[17] R. Schlüter *et al.*, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 649–652.