Karlsruhe Institute of Technology

Speech Feature Enhancement for Environments with Background Music

Studienarbeit
von

Evgeniy Shin

Institut für Anthropomatik
Karlsruher Institut für Technologie
Interactive Systems Laboratories

Betreuer: Florian Kraft, Prof. Alex Waibel

30. Januar 2012

Abstract

In this work an improvement for particle filter based speech feature enhancement for automatic speech recognition is proposed. This improvement is aimed to speech enhancement in environments with non stationary, voiced noises such as music. For this reason a voiced-unvoiced classification of distorted speech is performed on a frame-by-frame basis. A particle filter is served as noise tracking framework to estimate clean speech features with comparison of voicedness class priors, i.e. acoustic model of speech and autocorrelation based classification. Further comparison to unvoiced noise is performed.

Contents

1. Introduction

Noise reduction plays a big role in automatic speech recognition, because even a weak noise leads often to significant rising of word error rates.

But noise is an indefeasible part of every recording made in any real environment. For recordings, which are made outside, one can think about nature sounds, like wind, water, the birds or also traffic and industry noises. In closed environments there are a lot of noises too, e.g. heating, ventilation, air conditioning, office noises, steps, crowd etc.

Noise reduction shows to be a very difficult and challenging problem due to complex noise patterns and mixtures of different noises. Performance of automatic speech recognition (ASR) is affected by noise types differently. Voiced noise such as music affects performance of ASR systems stronger. One reason for that is the energy concentration of voiced noise in dedicated frequency bands, which leads to strong speech signal distortion in that bands.

1.1 Related work

A speech enhancement system generally consist of two major components: The estimation of speech signals, and the estimation of noise. The estimation of noise becomes a difficult problem if noise is assumed as non stationary i.e. changing over time.

Many solutions were proposed to deal with non stationary noises. Numerous evolved from audio source separation research. In many cases noise can be seen as just another source signal, that is not of interest for the given problem, e.g. background music or even another speech signal (cross talk). A review of modern source separation techniques can be found in [VJA+05]. Two main approaches can be distinguished here. First a source separation based on spatial diversity of sources e.g. [VHU10], which assumes polyaural recordings. In general there should be so many channels in the recording as there are sources. Second such assumptions are not needed, but a priori information about sources is exploited. This a priori information is generally presented as learned signal codebooks. Evaluation of several strategies single sensor source separation on speech and music sources is performed in [BRF08]. Those strategies can be generalized for application on other sources. In [BRF08] the non-stationarity of sources is not modeled explicitly, i.e. the model is parametrized with time variables, and this parameter is another dimension of the model, which is being estimated in a Maximum Likelihood (ML) sense with the other parameters.

In [FW06] Wölfel and Faubel have integrated the particle filter framework in an ASR system to track noise for speech feature enhancement. Their particle filter makes use of two

pass speech recognition systems. At the first pass the ASR system produces hypotheses. The particle filter models the speech signal with a static general Gaussian Mixture Model (GMM). The hypotheses are used in the second pass to estimate a phone-specific speech GMM. In [Wöl08] the authors have improved their approach. They have joined the particle filter with a multi-step linear prediction framework to also treat the reverberant distortions of speech signals.

The main drawback of phone-specific speech modeling, used in particle filter, is that the hypotheses, produced at the first pass do not have be correct. Errors in the hypotheses lead to choices of mismatched phone-specific model of speech. Using this model makes enhanced speech features tying to the wrong hypotheses. This effect is discussed in the next chapter.

An improvement addressed to this problem is offered by Heger in [Heg07]. Their idea is to use speech models based on phone classes. The classification of these classes is expected to be more robust to noise. The way of proceeding was the same with phone specific speech models: Voiced or unvoiced decisions for each frame were made according to the hypotheses from the first ASR pass, and than used in the second pass to model speech as a sequence of voiced or unvoiced specific GMMs.

In the work he studied such phone classes as: voiced-unvoiced, vowels-consonants and also data-driven clustered classes. Of the main interest were voiced-unvoiced classes, because of big differences in the nature of voiced and unvoiced signals, i.e. energy distribution over frequencies.

The weakness of this technique is poor robustness of the ASR system in case of non-stationary noises.

1.2 Goal

In this work we aim at enhancement of speech features robust to background music in context of automatic speech recognition. The non-stationarity of the noise is explicitly modeled and tracked within a particle filter framework, while for speech signals not being tracked some extra information about speech signal is used, i.e. a voiced-unvoiced classification of distorted speech performed on a frame-by-frame basis. The voiced-unvoiced classification is an a priori information for the particle filter. Therefore it is generated to be used as input of the particle filter framework.

In this work an improvement to the Heger's system [Heg07] is suggested. The idea is to use another technique to perform the voiced-unvoiced classification. This technique is based on improved autocorrelation algorithm, proposed by proposed Boersma in [Boe93] and implemented in Praat, a system for doing phonetics by computer[Boe02]. The algorithm will be described in one of the next sections.

Another point of interest in this work is to study the influence of different noise types on the performance of voiced-unvoiced classification and its effect on the performance of the particle filter. Experiments with two types of noise are performed: unvoiced noise e.g. wind, ocean, and voiced noise, e.g. music.

2. Particle filter

2.1 Particle filters in ASR

In this section the particle filter framework and its application in automatic speech recognition is introduced. Main steps in the particle filter processing chain are reviewed.

Particle filters, in statistics also known as sequential Monte Carlo Methods are usually used to estimate Bayesian models in which the latent variables are connected in a Markov chain – similar to HMMs but typically where the state space of the latent variables is continuous rather than discrete [Wik12]. It means that if we track a state of some dynamical system, which model is not necessary restricted to Gaussian distributions and is too complex to solve it analytically we can provide an optimal estimate of the model and track the state with help of the particle filter.

One application of the particle filter in automatic speech recognition is noise tracking. This idea is founded on the knowledge, that some noises can be quite well estimated locally and are of a periodic nature e.g. music, steps, street traffic etc.

Figure 2.1 depicts steps of the particle filter algorithm. y_k represents features for the frame k of the noise distorted signal, e.g. spectral coefficients. d_k is than the estimated noise spectrum of the frame k. S particles of the particle filter are S possible noise spectra of frame k. Particles are denoted by $d_k^{(s)}$, which refers to estimated noise spectrum for frame k by particle s, where $s = 0, \ldots, S$. The particles are evaluated with the evaluation function $\beta(d_k^{(s)}, y_k) \sim p(y_k|d_k^{(s)})$ i.e. the probability of noisy spectrum given a distortion spectrum.

Steps in the particle filter algorithm:

1. Initialization: Suppose S particles with some initial configuration. This particles are weighted samples of noise for a given frame. These weights are initially set accordingly to the uniform distribution.

2. Importance sampling: Pick S times one particle in respect to the particle weights i.e. particles with big weight are picked more often.

3. Particle evolution: At this step particles get their new configuration. An appropriate model is used to calculate the new possible noise spectra for each particle. Some rushing is also added in order to be able to track in case of deviation from the model.

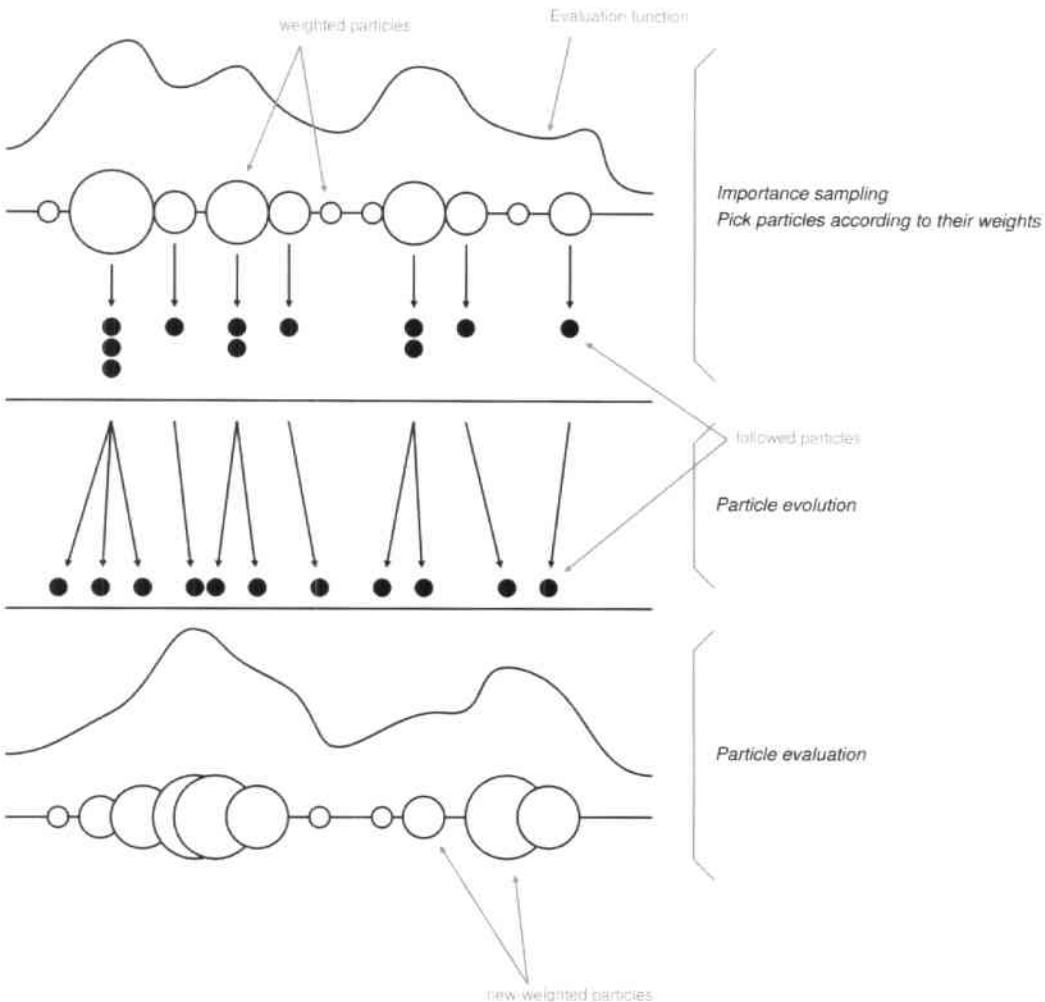

Figure 2.1: Main steps in the particle filter processing chain.

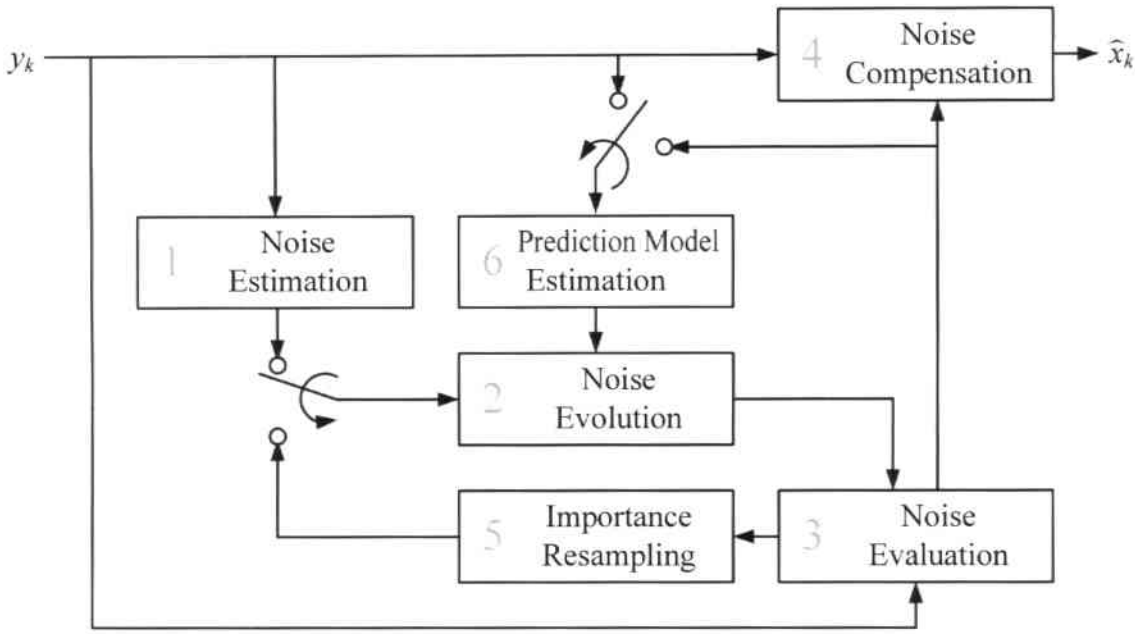

Figure 2.2: General flowchart of frame based speech feature enhancement for non-stationary additive distortion using a particle filter with importance resampling.

4. Particle evaluation: New weights of the particles are calculated with the evaluation function.

It is important to notice, that the model of clean speech plays the primary role in the evaluation step. The evaluation function, more exactly the probability $p(y_k|d_k^{(s)})$ is proportional to the probability of clean speech estimate by given distortion estimate $d_k^{(s)}$ referable to the particle s:

$$p(y_k|d_k^{(s)}) \sim p_{\mathcal{M}}(x_k|d_k^{(s)}),$$

where $\mathcal{M}$ is a speech model and x_k estimated speech feature for frame k.

2.2 Particle filter framework

The particle filter framework, that we use in this work is presented by Faubel and Wölfel in [Wöl08]. The authors present a generalized particle filter framework, which is capable of jointly tracking noise and reverberation on a frame-by-frame basis.

In this section we briefly restate their approach. Schematics of joint particle filter is given in Figure 2.2. The figure is taken from [FW06].

2.2.1 Reverberation Estimation

The revebration estimation is based on multi-step linear prediction (MSLP), which in contrast to the linear prediction (LP) aims to predict a signal after a given delay D, the so called step-size. With the prediction error e[n] MSLP can be formalized as:

$$y[n] = \sum_{m=1}^{M} c_m y[n-m-D] + e[n],$$

where $c_1, c_2, \ldots, c_M$ present the LP coefficients, $y[n]$ the observed signal and M the model order. To solve for the MSLP coefficients $\mathbf{c} = [c_1, c_2, \ldots, c_M]^T$ the minimum square error estimate is used.

An estimate of the reflection sequence $r[n]$ is obtained by filtering the observation sequence $y[n]$ with the MSLP filter

$$r[n] = \sum_{m=1}^{M} c_m y[n - m + D],$$

The reflection sequence $r[n]$ is now converted into short-time power spectra $\mathbf{r}_k$. This highly non-stationary distortion estimate is treated just like an additive distortion and removed from the distorted sequence $y[k]$ by spectral subtraction. The distorted $y[k]$ is respectively estimated for all frames, $k = 0, \ldots, K$, from $y[n]$.

2.2.2 Distortion Estimation and Particles Initialization

The first step is particle initialization by drawing samples from the prior particle density. The prior density of additive distortion $p(\mathbf{a}_0)$ is estimated indirectly via overall distortion density $p(\mathbf{d}_0)$ and estimated reverberation distortion density $p(\mathbf{r}_0)$:

$$p(\mathbf{a}_0) = \ln(\exp p(\mathbf{d}_0)) - \exp p(\mathbf{r}_0)),$$

where $p(\mathbf{d}_0)$ derived on silent regions of the input signal which contains additive and convolutive distortions and $p(\mathbf{r}_0)$ is estimated over all frames derived on the reflection sequence $\mathbf{r}_k$.

2.2.3 Particle evolution

The evolution for each particle $\mathbf{p}_k^{(s)}$, $s = 0, \ldots, S-1$, is estimated by an autoregressive process with AR-matrix $\mathbf{P}_{k-1}$ for each frame k:

$$\mathbf{p}_k^{(s)} = \mathbf{P}_{k-1}\mathbf{p}_{k-1}^{(s)} + \epsilon_k^{(s)}; \qquad \epsilon_k^{(s)} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

The AR-matrix is given by

$$\mathbf{P}_k = \left[\begin{array}{c|c} \mathbf{A}_k & 0 \\ \hline 0 & \mathbf{S}_k \end{array}\right],$$

where $\mathbf{A}_k$ is the additive distortion matrix and $\mathbf{S}_k$ is the matrix of scale terms for the reverberation distortion. Dimension of $\mathbf{S}_k$ is 1 for *shared scaling factor* and B for *individual scaling factor for each frequency bin*, where B is the number of frequency bins.

2.2.4 Distortion combination

Now, the distortion samples $\mathbf{d}_k^{(s)}[b]$ are calculated for each particle $\mathbf{p}_k^{(s)}$ and frequency bin b as

$$\mathbf{d}_k^{(s)}[b] = \ln\left\{\exp \mathbf{a}_k^{(s)}[b] + \mathbf{s}_k^{(s)}[b] \exp \mathbf{r}_k[b]\right\},$$

where $a[b]$ represents additive distortions, $r[b]$ represents the spectral distortion due to reverberation and $s[b]$ represents the scale terms for $r[b]$.

2.2.5 Distortion evaluation

With the prior speech density $p_{speech}(\cdot)$ each distortion sample $\mathbf{d}_k^{(s)}$ is evaluated according to the likelihood

$$p(\mathbf{y}_k|\mathbf{d}_k^{(s)}) \sim p_{speech}(\mathbf{y}_k + \ln(1 - e^{\mathbf{d}_k^{(s)} - \mathbf{y}_k})) \tag{2.1}$$

and its normalized weight

$$w_k^{(s)} = \frac{p(\mathbf{y}_k|\mathbf{d}_k^{(s)})}{\sum_{m=1}^{S} p(\mathbf{y}_k|\mathbf{d}_k^{(m)})}$$

is calculated. Thereby $1 - e^{\mathbf{d}_k^{(s)} - \mathbf{y}_k}$ is a clean speech estimation, given signal $\mathbf{y}_k$ and currently evaluated distortion $\mathbf{d}_k^{(s)}$.

2.2.6 Distortion compensation

The clean feature is estimated with the distortion samples $\mathbf{d}_k^{(s)}$ and their corresponding importance weights $w_k^{(s)}$ over all particle samples S using the non-linear relationship between $\mathbf{x}_k$, $\mathbf{d}_k$ and $\mathbf{y}_k$ as in

$$\mathbf{E}\left\{\mathbf{x}_k|\mathbf{y}_{(1:k)}\right\} = \sum_{s}^{S} w_k^{(s)} \left(\mathbf{y}_k + \ln(1 - e^{\mathbf{d}_k^{(s)} - \mathbf{y}_k}\right)$$

2.2.7 Importance Resampling and Prediction Model Estimation

This is the last, but very important step in the particle filter processing chain. After every time step, the particles are resampled, in order to avoid the concentration of the vast majority of probability mass in very few particles. The prediction model is also new estimated by the dynamic autoregressive process.

3. Voiced-unvoiced classification

The tying of the phone-specific speech model to the ASR hypotheses, which the model is based on, makes the application of the pnone-specific particle filter not efficient by high error rates, which is the case in very noisy environments. This effect can be explained as following: the particles, which improve the probability of clean speech estimate, are weighted higher, even if the clean speech is estimated wrongly (an ASR hypothesis is not correct). The phone-specific particle filter performs expected good, if clean speech features are estimated on the conversation reference.

To increase performance of speech feature enhancement by the particle filter the advantage of voiced-unvoiced classification of the distorted signal on a frame-by-frame basis is taken. The motivation is more robust voiced-unvoiced speech.

For this classification we use two methods and compare them. The first is an autocorrelation-based method, proposed by Boersma in [Boe93] and implemented in Praat, a system for doing phonetics by computer [Boe02]. We compare this method to the hypoteses based method, suggested by Heger in [Heg07].

In order to evaluate the performance of these techniques references of clean speech signals are being used. In the following it is denoted as reference based classification.

In the next sections we introduce these techniques.

3.1 Hypotheses based method

The hypotheses based method is presented in [Heg07]. The idea is use an automatic speech recognition system to produce hypotheses and predefined phone classes for the classification.

These phone classes are voiced and unvoiced classes for our application. The voiced class is a set of phones, which are a priori known as voiced. The unvoiced class is respectively a set of phones, known as unvoiced.

The speech recognizer computes a hypotheses transcription of speech on distorted signal. This transcription is a sequence of phones. All we need is to check out, which of them belong to voiced or unvoiced phone classes.

One disadvantage of the hypotheses-based method is that the voiced-unvoiced specific model cannot be used in the first pass, because no voiced-unvoiced classification is done yet.

Another disadvantage of this method, as it was mentioned before, is a big number of errors in hypotheses in case of strong signal distortion. Therefore a more robust method is suggested in the next section.

3.2 References based method

The idea of this method is the same as for the hypotheses based method. The only difference is that here the forced (with reference) Viterbi-alignment on HMM is used to produce a sequence of phones. Every phone is than classified with the same predefined classes of phones.

It is important to notice here, that the result of the voiced-unvoiced classification with this method differs by different noise types and signal-to-noise ratio. The reason for it is that the Viterbi algorithm finds the best alignment in according to the acoustic information.

3.3 Autocorrelation based method

To make a decision, if one frame is voiced or unvoiced a presence of speaker's pitch is used, as the voiced speech is known as quasi-periodic, consisting of a fundamental frequency corresponding to the pitch and its harmonics. It means if we could hear or see that there is a fundamental frequency in the frame spectrum, than we decide for voiced speech. In another case we decide for unvoiced speech, which is stochastic in nature and do not consist of a periodic part in the spectrum. It can also be modeled as white noise. Therefore an autocorrelation-based method for pitch detection is used.

Such a method presented by Boersma in [Boe93] is straightforward and robust for periodicity detection. It is working in autocorrelation domain.

"By definition, the best candidate for the acoustic pitch period of a sound can be found from the position of the maximum of the autocorrelation function of the sound, while the degree of periodicity (the harmonics-to-noise ratio) of the sound can be found from the relative height of this maximum. However, sampling and windowing cause problems in accurately determining the position and height of the maximum. These problems have led to inaccurate time-domain and cepstral methods for pitch detection, ..."[Boe93]

The windowing problem is present almost in every short-term analysis of periodic signals. As it's name says, the application of a window function leads to a signal distortion on the rim of the window. The reason is a window shape. The problem is solved in [Boe93]. The authors propose to divide the autocorrelation of the windowed signal by the autocorrelation of the window.

Next problem is the sampling problem. The reason is that the autocorrelation computed from the sampled signal is also a sampled function. It means, that the position of peaks is determined with a precision of sampling interval, which implies not very accurate estimation of the pitch period. To solce this problem an interpolation of the autocorrelation function is proposed [Boe93].

A summary of the algorithm is give here:

1. Preprocessing: Upsampling is performed to remove the sidelob of the Fourier transform of the window for signal components near Nyquist frequency.
2. Compute the global absolute peak value of the signal.
3. Short-term analysis
    a) Perform signal segmentation and windowing. The length of this segment (the window length) is determined by a parameter.

b) Perform a Fast Fourier Transform.

c) Square samples in the frequency domain.

d) Perform a Fast Fourier Transform, which gives a sampled version of the autocorrelation function of the windowed signal.

e) Divide by the autocorrelation of the window. This gives a sampled version of autocorrelation function estimate of the original signal.

f) Compute the continuous version of autocorrelation function with help of interpolation.

g) Evaluate pitch candidates. The strength of unvoiced candidate is computed with voicing and silence thresholds. These depend on the global absolute peak value of the signal. If there are no autocorrelation peaks above the voicing threshold and less than silence threshold, a frame bears a good chance of being analyzed as voiceless. The maximum of the autocorrelation function for the current frame gives the actual pitch value.

4. Find the optimal path through the pitch candidates with the aid of dynamic programming by optimizing the cost associated with every possible path. This cost is computed with help of two transition cost parameters: *VoicedUnvoicedCost* and *OctaveJumpCost*.

One disadvantage of this algorithm is that it rather aims to accurate pitch estimation than to voiced-unvoiced classification. On the other side and in contrast to hypotheses-based classification this algorithm can be applied as early as needed, because the classification happens prior to the speech enhancement.

This algorithm aims to estimate the pitch in the frequency band between *minimum pitch* and *maximum pitch*. The default values are: 75Hz for the *minimum pitch* and 600Hz for the *maximum pitch*. Therefore if noises do not distort this frequency band the algorithm is robust at high SNRs. Otherwise noise obscures the underlying fundamental frequency. Unvoiced noise should not strongly affect the voicing and silence thresholds, because they depend on the global absolute peak value of the signal. The transition costs: *VoicedUnvoicedCost* and *OctaveJumpCost* discriminate pitch paths with many voiced-unvoiced transitions and rash changes of pitch frequency, which helps to resist the action of voiced noise. These all explain the robustness of the autocorrelation-based voiced-unvoiced classification.

4. Experiments

In this section results of the experiments are presented. But first the speech recognition toolkit, datasets and acoustic models of speech used in the particle filter are discussed.

4.1 Setup

4.1.1 Automatic speech recognition toolkit

The automatic speech recognition engine that we use in this work is JRTk[M+04], which is developed and maintained by the Interactive Systems Laboratories at two sites: Karlsruhe Institut of Technology, Germany and Carnegie Mellon University, USA.

The speech recognition system setup used in this work processes the signal utterance by utterance i.e. presegmented utterances are being processed one after another.

The particle filter, that was used in the following experiments, is discussed in chapter 2. It is also integrated in JRTk.

4.1.2 Particle filter and acoustic models

The particle filter is configured for our experiments to use 100 particles. It is the default option in this particle filter framework. Another important option is the *transcription* option. If the transcription is given, the particle filter tries to use specific speech models. Using this option and predefined speech models makes it able to work with different class-based models.

Two voiced-unvoiced specific acoustic models were integrated in the particle filter framework within equation 2.1. During particle processing the particle filter receives a special transcription, which describes each frame, if it is voiced or unvoiced. In respect to this decision a specific speech model is chosen. If there is no decision for any frame in the transcription a general speech model is applied. This model does not account for differences between voiced and unvoiced speech, but helps to estimate the speech signal, in case when voiced-unvoiced classification fails. In our experiments and for the test data used in the experiments this case has not occurred.

To train voiced specific, unvoiced specific and general speech models the sampling material from [FW06] was used: "... train acoustic model on the close talking channel of meeting corpora and merge it with the Translanguage English Database (TED) corpus summing

up to a total of approximately 100 hours of training material. The speech data was sampled at 16 kHz. Speech frames were calculated using a 10 ms Hamming window. For each frame, 13 mel frequency cepstral coefficients or warped MVDR cepstral coefficients were obtained through a discrete cosine transform from the Fourier transformation or the warped MVDR spectral envelope. Thereafter, linear discriminant analysis was used to reduce the utterance based cepstral mean normalized features..."[FW06].

Using the merge-and-split training procedure speech GMMs, containing 64 Gaussians for each model, were trained.

The samples were split into two sets respectively to the two classes of phones:

$$\begin{aligned} \mathcal{C}_{voiced} &= \{B, D, G, JH, V, DH, Z, ZH, M, N, NG, W, R, Y, L, ER, \\ &\quad AY, OY, EY, IY, AW, OW, EH, IH, AO, AE, AA, \\ &\quad AH, UW, UH, AXR, IX, AX, XL, XM, XN\} \end{aligned} \tag{4.1}$$

$$\mathcal{C}_{unvoiced} = \{P, F, TH, T, S, SH, HH, CH, K\} \tag{4.2}$$

For the general speech model phone samples of both classes have been processed at once.

To make the direct comparison of voiced-unvoiced and general speech particle filter possible, the voiced-unvoiced models were trained on the same samples as the general speech model, i.e. the number of samples used for each voiced and unvoiced models was smaller than for the general speech model.

The resulting models are represented with mixtures of 64 Gaussian models. Mixtures with 32 Gaussians for voiced-unvoiced specific models have also been tested to take in account smaller number of samples used to estimate voiced and unvoiced models.

The noise model of the particle filter is being updated on the fly. A speech activity detector determines, if there are enough noise frames in a current utterance. In such case the noise model is being updated.

4.1.3 Dataset

For the evaluation a close talking channel of lecture speech with 16kHz and 40 minutes duration was chosen. This data was mixed with noise with 10 and 20dB signal-to-noise ratio (SNR). Two different noise types were used: ocean waves and wind as unvoiced noise and music without vocal as voiced noise.

Ocean wave sounds were taken from [gra07]. "Ocean waves recorded from about 250 feet up. Recorded on a Zoom H4 with the mics covered in zeppelin fur at 96kHz 24-bit and downsampled to 48kHz. Because of the distance, the sound is as much about white noise as waves." [gra07] This sound was downsampled to 16kHz to mix with speech data.

The music dataset was provided by authors of [RLP+06]. This dataset consists of 600 pop, jazz and classical compositions in karaoke and midi formats. We did not distinct between pop, jazz and classical music, therefore compositions from all sets were randomly selected. The tracks were converted to PCM waveform data with 16kHz sample rate. In order to mix data in proper SNR the root mean square (RMS) amplitude of all tracks was normalized to the RMS amplitude of speech signal. The selected tracks were concatenated and mixed with speech data.

Mixtures with 3 signal-to-noise ratios were created for the experiments: clean speech (nomix), 20 dB and 10 dB.

4.2 Systems

For the experiments 50 different system configurations were tested. These are systems with different voiced-unvoiced classification methods, systems with only general speech particle filter and systems with no particle filter enabled; each of them with two noise types and with three SNRs: clean speech, 20dB and 10dB.

As a front-end a warped minimum variance distortionless response (MVDR) [WM05] envelope was used. Signal frames were calculated using 16 ms Hamming window. A filterbank with 129 bins was applied. In order to prevent the particle filter working in high dimensional space, the logarithmic spectral domain, after cepstral to 20 dimensions in the cepstral domain, is used.

Every test consists of three steps:

1. First pass recognition: In this step the ASR system computes hypotheses-transcriptions for the conversations in signal.
2. Speaker adaptation. Hypotheses from the first recognition are used for Maximum Likelihood Linear Regression (MLLR) and feature space MLLR. Vocal tract length normalization (VTLN) is also applied.
3. Second pass. The ASR system recomputes new hypotheses.

The voiced-unvoiced classification occurs before the first pass for systems with the autocorrelation-based classifier. Therefor a praat [Boe02] script is called. It converts an estimated by praat pitch sequence into voiced-unvoiced transcription, that is understood by the particle filter framework. For systems with hypotheses and references-based classification, the voiced-unvoiced transcription is generated just before an adaptation step. Hypotheses from the first pass force the Viterbi alignment. Sequence of phones is than converted in the voiced-unvoiced transcription with help of the predefined classes: voiced 4.1 and unvoiced 4.2.

Here is an overview of tested system classes:

- no-pf: Systems with no particle filter enabled.
- *gs-pf*: Systems with the general-speech particle filter. To model the speech signal, only the general speech model was used here. The particle filter was applied in both first and second passes, so as by adaptation.
- *ref-vuv*: Systems with the voice-unvoiced specific particle filter. Voiced-unvoiced classification bases on reference from ASR. In the first pass the general speech particle filter was applied. In the adaptation step and in the second pass the voiced-unvoiced particle filter was used. Thereby for the voiced-unvoiced classification a forced Viterbi alignment on the ASR reference was utilized.
- *hypo-vuv*: Systems with the voice-unvoiced specific particle filter. Voiced-unvoiced classification bases on hypotheses from ASR. As in *ref-vuv* systems the general speech particle filter was applied in the first pass; in adaptation and second pass a forced Viterbi alignment on the hypotheses from the first pass was performed.
- *autocor-vuv*: Systems with the voice-unvoiced specific particle filter. Voiced-unvoiced classification bases on autocorrelation method implemented in praat [Boe02]. Praat version 5.2.44 (23 September 2011) was used in tests. The voiced-unvoiced specific particle filter was applied in both first and second passes and in adaptation.

Furthermore some extra system configurations were tested. These are *autocor-vuv-32*, *autocor-vuv-sp* and *autocor-vuv-sp-32* systems. *autocor-vuv-32* are *autocor-vuv* system configurations, where the voiced-unvoiced speech model were represented by 32 Gaussians

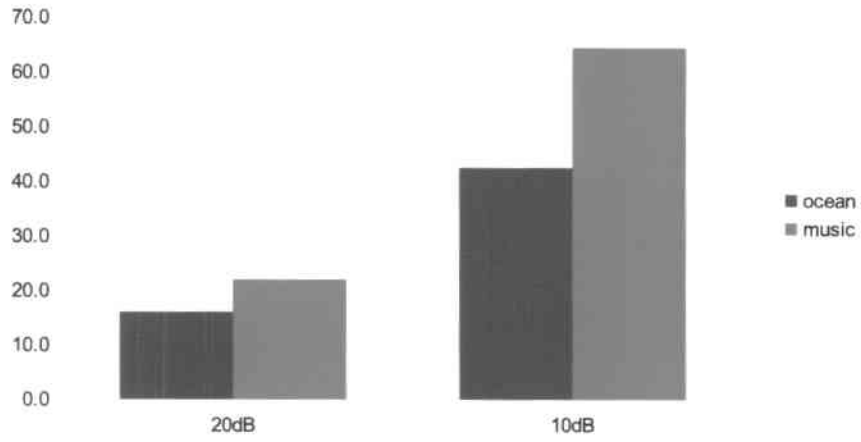

Figure 4.1: WER for music and ocean noises of systems with no particle filter.

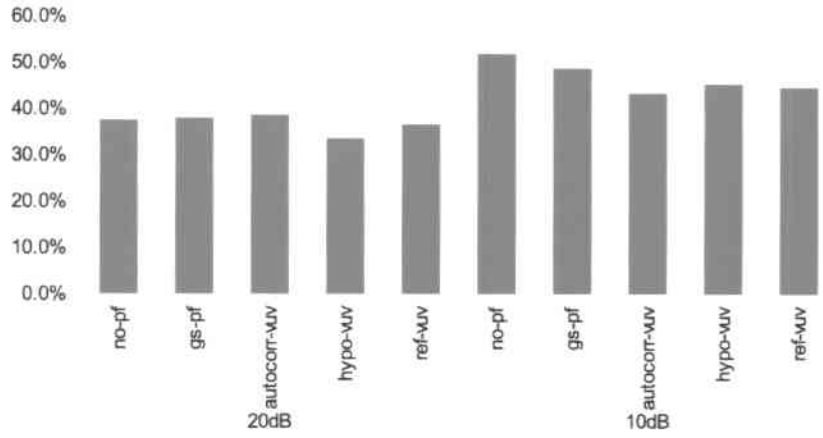

Figure 4.2: Relative growth of WER in case of music noise in comparison to ocean noise.

instead of 64. In *autocor-vuv-sp* the voiced-unvoiced specific particle filter was used only in adaption and second pass. In the first pass the general speech particle filter was applied. *autocor-vuv-sp-32* systems are equal to *autocor-vuv-sp* systems except voiced-unvoiced speech models, which are represented each by 32 Gaussians in GMMs.

4.3 ASR and noises

Results of the experiments for two noise types: music and ocean (unvoiced) noise show, that the speech recognizer performance is impacted differently by these two noise types. The growth of word error rate (WER) in case of music noise on the ASR system without particle filter is 6 points at 20dB SNR and more than 22 points at 10dB SNR as shown in figure 4.1. The WERs are from the second pass recognition. Voiced noise (music) seems to disturb the speech signal much more as unvoiced (ocean).

A relative growth of WERs in case of voiced noise in percent of WERs in case of unvoiced noise is shown in the figure 4.2. With no particle filter enabled the background music leads to ASR performances at about 30 percent worse as in case of ocean noise within the same SNR. An application of the particle filter does not improve this relation.

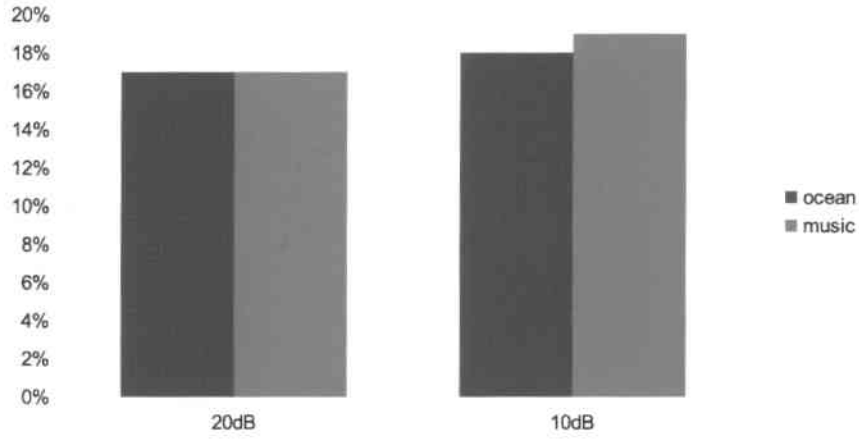

Figure 4.3: Error rates of autocorrelation-based classifier on ocean and music noises.

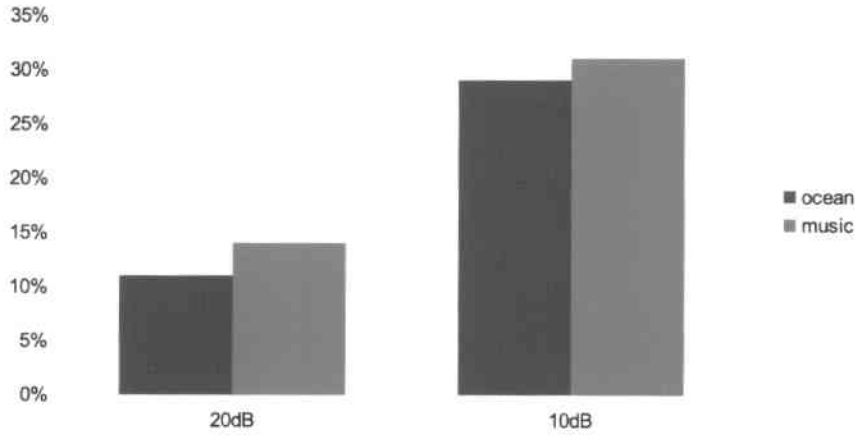

Figure 4.4: Error rates of hypotheses-based classifier on ocean and music noises.

4.4 Voiced-unvoiced classifiers

In this section the performances of autocorrelation- and hypotheses-based voiced-unvoiced classifiers are compared. More exactly classification error rates, i.e. an average error pro 100 frames of autocorrelation- and hypotheses-based methods are compared to each other. As a reference classification results of reference-based method for clean speech signal were taken. As it was mentioned in section 3.2, this is not an optimal criterion, but the only way within out setup to compare the classification performances data-driven.

As it shown in the figure 4.3 the voiced-unvoiced classification error rates of the autocorrelation-based algorithm at 20dB SNR for both music and ocean noise are the same. At 10dB it performs better in case of unvoiced noise. The hypotheses-based method has higher error rates (figure 4.4) for voiced noise at both 20 and 10dB SNR.

The figures 4.5 and 4.6 state that the hypotheses-based method gives the autocorrelation-based method no chance if a clean speech signal is to classify. At 20dB distortion the error rates are comparable, but hypotheses-based method is still a bit better. At 10dB the autocorrelation-based classifier becomes much better. Furthermore the performance of autocorrelation-based classifier is weak affected by growing SNR, which speaks for ro-

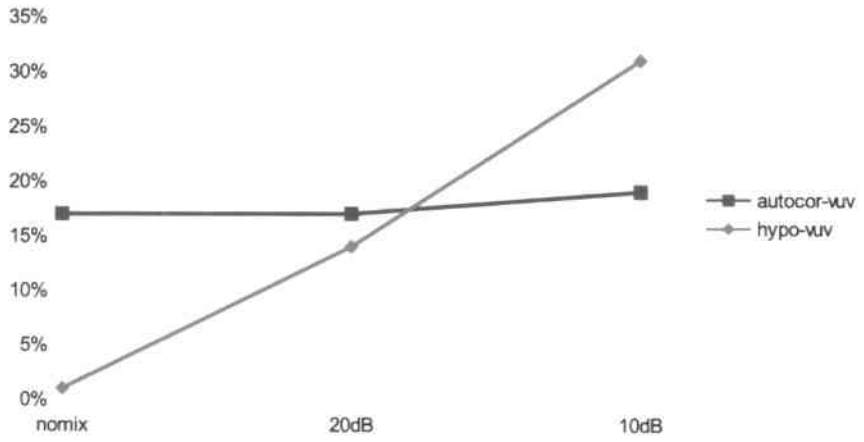

Figure 4.5: Voiced-unvoiced classification error ratio on voiced (music) noise.

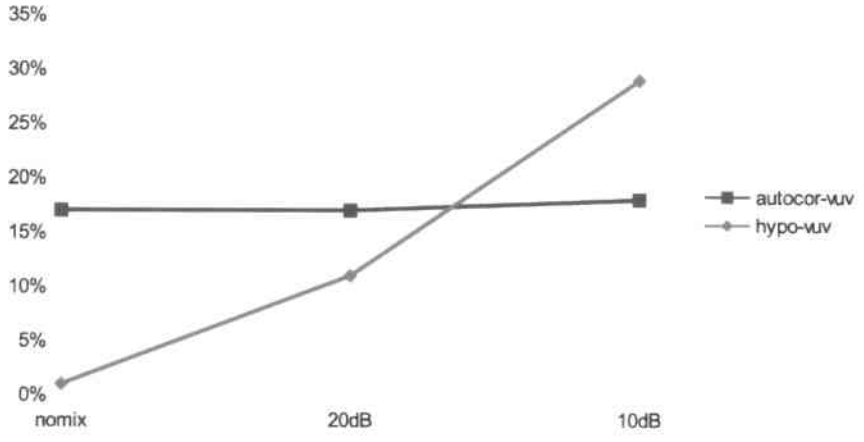

Figure 4.6: Voiced-unvoiced classification error ratio on unvoiced (ocean) noise.

bustness of the algorithm. Performances of voiced-unvoiced classifiers by stronger noise is not tested, because of extremely high, not acceptable word error rates of ASR hypotheses in this case.

4.5 Voiced-unvoiced specific particle filters

The table 4.1 shows the results of the experiments on ASR systems with different types of particle filters.

The results of the tests on *ref-vuv* systems are against expectation not always the best and in several tests even worse as for *hypo-vuv* systems. It is the case by the clean speech test, without any additive distortion. The *hypo-vuv* system performs better as the *ref-vuv*. A reason for it may be a mismatch between references and acoustic information in the signal e.g. speaker pronounces some words in a different manner as in a dictionary. By 20dB SNR the both systems perform equal. Because the mismatch by *ref-vuv* system is compensated by higher error rate of hypotheses-based voiced-unvoiced classifier. In case of 10dB background music the *ref-vuv* system performs better.

As for *autocor-vuv* systems their results were as expected. According to the performance of autocorrelation-based voiced-unvoiced classifier, *autocor-vuv* systems are better only in case of strong signal distortion, namely 10dB both by voiced and unvoiced noises. Their gain in comparison to *hypo-vuv* systems is not high: about 2.8% in case of music noise and about 1.5% in case of ocean noise.

Comparison of the best voiced-unvoiced particle filter systems with general speech particle filter system gives clear results: systems with voiced-unvoiced specific particle filter have performed a bit better in all tests. The maximal gain is about 4% in case of 10dB background music.

The test of extra configurations: *autocor-vuv-32*, *autocor-vuv-sp* and *autocor-vuv-sp-32* does not improve on top of previous experiments.

	nomix		music				ocean			
			20dB		10dB		20dB		10dB	
	unadp	adp	unadp	adp	unadp	adp	unadp	adp	unadp	adp
no-pf	11.2	9.2	40.6	22.0	71.3	64.5	29.7	16.0	66.2	42.5
gs-pf	10.9	8.9	38.4	21.1	68.8	61.1	29.2	15.3	64.4	41.1
autocor-vuv	11.4	9.1	37.6	21.2	68.6	58.6	29.0	15.3	62.6	40.9
hypo-vuv	10.9	8.7	38.4	20.3	68.8	60.3	29.2	15.2	64.4	41.5
ref-vuv	10.9	9.0	38.4	20.8	68.8	60.8	29.2	15.4	64.4	40.6
autocor-vuv-32	11.7	8.9	37.9	20.9	68.7	59.3	29.5	16.0	63.1	42.1
autocor-vuv-sp	10.9	9.1	38.4	21.6	68.8	60.4	29.2	15.3	64.4	40.6
autocor-vuv-sp-32	11.5	9.4	38.8	20.6	68.2	60.3	28.6	15.5	63.9	42.7

Table 4.1: Main results. Evaluation measure used is WER. The systems are described in section 4.2

5. Conclusion

In this work the following goals were examined: first it was tested how different types of noise affect the performance of particle filter based speech feature enhancement and second if voiced-unvoiced classification can improve the performance of the standard particle filter approach especially in case of background music.

The experiments show, that the ASR performance using particle filter speech feature enhancement is more affected by voiced noise than by unvoiced. This can be explained by different noise signal energy distribution on frequencies. At same signal-to-noise ratios, the energy of voiced noise is concentrated in compact regions, which leads to stronger distortion of speech signal. In contrast the energy of unvoiced noised is distributed over all frequencies.

According to the experiments voiced-unvoiced specific particle filter slightly improves ASR performance in case of voiced noise and brings fast no improvement in case of unvoiced noise. The comparison of autocorrelation-based voiced-unvoiced classification shows better results than hypotheses-based classification at 10dB signal-to-noise ratio. This is also confirmed by experiments that show classification error ratios. These results can be explained by greater robustness of autocorrelation-based algorithm. The results after the first pass of ASR help to interpret the improvements by autocorrelation-based methods. After the first pass autocorrelation-based methods have better hypotheses, which leads to better signal adaptation; the improvement propagates to the second pass.

Interesting results were obtained with reference-based method. They are not much better and sometimes even worse as the results with hypotheses-based method. One guess is that the Viterbi alignment forced by references is fooled by strong background noise more than if it forced by hypotheses, because hypotheses are estimated on the distorted signal. Of course errors in the experiment setup were also possible.

The experiments show that the particle filter with general speech acoustic model used for speech feature enhancement improves the performance of the ASR noticeably. However using class-based models for speech, such as voiced-unvoiced or phone-specific brings no significant improvement. A better performance of voiced-unvoiced classifier would eventually not increase the performance of the particle filter.

Bibliography

[Boe93] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, vol. 17. Amsterdam, 1993, pp. 97–110.

[Boe02] ——, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[BRF08] R. Blouet, G. Rapaport, and C. Fevotte, "Evaluation of several strategies for single sensor speech/music separation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 37–40.

[FW06] F. Faubel and M. Wölfel, "Coupling particle filters with automatic speech recognition for speech feature enhancement," in *Ninth International Conference on Spoken Language Processing*, 2006.

[gra07] graysound, "Ocean from 250 feet jmausti," 2007, [Online; accessed 6-January-2012]. [Online]. Available: http://www.freesound.org/people/greysound/sounds/31435/

[Heg07] D. Heger, "Speech feature enhancement using particle filters with class-based phoneme models," 2007.

[M+04] F. Metze *et al.*, "Jrtk online documentation," 2004.

[RLP+06] D. Rizo, P. León, J. Pedro, C. Pérez-Sancho, A. Pertusa, and J. Iñesta, "A pattern recognition approach for melody track selection in midi files," *Distribution*, pp. 61–66, 2006.

[VHU10] D. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 241–244.

[VJA+05] E. Vincent, M. Jafari, S. Abdallah, M. Plumbley, and M. Davies, "Blind audio source separation," *Queen Mary, University of London, Tech Report C4DM-TR-05-01*, 2005.

[Wik12] Wikipedia, "Particle filter — Wikipedia, the free encyclopedia," 2012, [Online; accessed 4-January-2012]. [Online]. Available: http://en.wikipedia.org/wiki/Particle_filter

[WM05] M. Wolfel and J. McDonough, "Minimum variance distortionless response spectral estimation," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 117–126, 2005.

[Wöl08] M. Wölfel, "A joint particle filter and multi-step linear prediction framework to provide enhanced speech features prior to automatic recognition," in *Hands-Free*