Coupling Particle Filters with Automatic Speech Recognition for Speech Feature Enhancement

Friedrich Faubel and Matthias Wölfel

Institut für Theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
wolfel@ira.uka.de

10.21437/Interspeech.2006-10

Abstract

This paper addresses robust speech feature extraction in combination with statistical speech feature enhancement and couples the particle filter to the speech recognition hypotheses.

To extract noise robust features the Fourier transformation is replaced by the warped and scaled minimum variance distortionless response spectral envelope. To enhance the features, particle filtering has been used. Further, we show that the robust extraction and statistical enhancement can be combined to good effect.

One of the critical aspects in particle filter design is the particle weight calculation which is traditionally based on a general, time independent speech model approximated by a Gaussian mixture distribution. We replace this general, time independent speech model by time- and phoneme-specific models. The knowledge of the phonemes to be used is obtained by the hypothesis of a speech recognition system, therefore establishing a coupling between the particle filter and the speech recognition system which have been treated as independent components in the past.

Index Terms: particle filters, automatic speech recognition, speech feature enhancement, phoneme-specific

1. Introduction

Particle filters (PF)s, a.k.a. sequential Monthe Carlo methods, originally developed for typical tracking applications like pursuing airplanes in radars [1], or persons in video images [2], are increasingly pervading other fields of engineering covering navigation, robotics, communications and (industrial) process control. Recently, they have found their way into speech recognition [3, 4] where they are used for the enhancement of speech features corrupted by noise. The advantage over classical methods like *spectral subtraction* [5] or *Wiener filtering* is that the PF allows the noise to be non-stationary.

The two critical aspects in PF design are the choice of the importance or proposal density and the particle weight calculation. A variety of different particle filter variants have been evaluated for the enhancement of speech features: auxiliary and likelihood PFs [6] as well as PFs with an extended Kalman filter proposal density [4]. In those approaches, however, the particle weight calculations were always based on a general, time-independent speech model approximated by a Gaussian mixture distribution. We propose to calculate the particle weights with a speech model that accounts for the dynamics of speech: a time- and phoneme-specific speech model, where the phoneme hypotheses stem either from forced alignment given the transcripts (a Wizard of Oz experiment to give an lower bound in terms of word error) or from a previous speech recognition pass.

2. Particle Filter Based Speech Feature Enhancement

To our best knowledge, Singh and Raj [7] were the first to use PFs in the context of speech feature enhancement for speech recognition. In their approach, a PF is employed to track the noise sequence that corrupts the speech signal. The estimated noise sequence is then used to clean or enhance the speech features. This is performed in spectral domain between two typical processing steps of speech feature extraction: after the filterbank which reduces the dimension of the input vector in the logarithmic mel power domain and before the transformation into the cepstral domain by a discrete cosine transformation. We briefly restate Singh's and Raj's approach in the following two sections (2.1, 2.2). Section 2.3 addresses the problem of divergence in conjunction with continuous speech recognition. In section 2.4 we increase noise robustness by replacing the Fourier transformation with a spectral envelope. Section 3 will finally introduce the phoneme-dependent speech model and explains how it can be coupled with the speech recognizer.

2.1. Tracking the Noise

For the PF to be applicable it is necessary to develop a *dynamical system model* (DSM) for the noise. Raj et al. proposed a 1st-order autoregressive model

$$n_t = A \cdot n_{t-1} + \varepsilon_t$$

where A is the transition matrix that is learned for a specific type of noise and n_t denotes the noise spectrum at time t. The ε_t terms are considered to be i.i.d. zero mean Gaussian, i.e. $\varepsilon_t \sim \mathcal{N}(0, \Sigma_{noise})$. Throughout this paper,

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma (x-\mu)}$$

shall denote a Gaussian distribution with mean μ and diagonal covariance matrix Σ. Using this notation the noise transition probability $p(n_{t+1}|n_t)$ can be written as

$$p(n_{t+1}|n_t) = \mathcal{N}(n_{t+1}; A \cdot n_t, \Sigma_{noise})$$

Denoting corrupted, clean and estimated clean speech spectra by y, x and $\tilde{x}$ respectively, the particle filtering stage can be outlined as follows:

1. At time zero ($t = 0$) noise hypotheses or particles $n_0^{(j)}$ ($j = 1, ..., N$) are *sampled* from the probability distribution $p(n)$ learned for noise spectra. Sampling from a probability distribution means simulating values (samples) of that probability distribution.

2. Modeling the probability distribution of clean speech spectra as a Gaussian mixture distribution

$$p(x) = \sum_{k=1}^{K} c_k \mathcal{N}(x; \mu_k, \Sigma_k) \quad (1)$$

the likelihood $l(n_t^{(j)}; y_t) = p(y_t|n_t^{(j)})$ of each noise hypothesis $n_t^{(j)}$ can be evaluated as

$$p(\tilde{x}_t^{(j)}) = \sum_{k=1}^{K} \frac{c_k \mathcal{N}(y_t + \log(1 - e^{n_t^{(j)} - y_t}); \mu_k, \Sigma_k)}{\left|1 - e^{\hat{n}_t^{(j)} - y_t}\right|}$$

where $\tilde{x}_t^{(j)}$ is the *imputed* clean speech spectrum which can be calculated from $n_t^{(j)}$ and y_t

$$\tilde{x}_t^{(j)} = y_t + \log(1 - e^{n_t^{(j)} - y_t}) \quad (2)$$

if the phase is discarded [6].

3. The normalized likelihoods or weights

$$\omega_t^{(j)} = \frac{p(y_t|n_t^{(j)})}{\sum_{m=1}^{N} p(y_t|n_t^{(m)})}$$

are now used to resample among the noise hypotheses $n_t^{(j)}$ $(j = 1, ..., N)$ which can be regarded as a pruning step where likely hypotheses are multiplied, unlikely ones are removed from the population.

4. Finally, the resampled noise hypotheses $n_t^{(j)}$ $(j = 1, ..., N)$ are used to generate new hypotheses for time $t + 1$ by sampling $n_{t+1}^{(j)}$ from the transition probability $p(n_{t+1}|n_t^{(j)})$, $j = 1, ..., N$.

Steps (2-4) are repeated with $t \mapsto (t + 1)$ until all time-frames of the speech data are processed.

2.2. Compensating for the Estimated Noise

Given a noise hypothesis $n_t^{(j)}$, the corresponding clean speech spectrum can be approximated by the *minimum mean square error* (MMSE) estimation [8, 7]

$$\hat{x}_t^{(j)} = y_t - \sum_{k=1}^{K} p(k|y_t, n_t^{(j)}) log(1 + e^{n_t^{(j)} - \mu_k})$$

where $p(k|y_t, n_t^{(j)}) = p(k|\tilde{x}_t^{(j)})$ is the normalized activity of the kth Gaussian in the Gaussian mixture distribution of clean speech

$$p(k|\tilde{x}_t^{(j)}) = \frac{c_k \mathcal{N}(\tilde{x}_t^{(j)}; \mu_k, \Sigma_k)}{\sum_{l=1}^{K} c_l \mathcal{N}(\tilde{x}_t^{(j)}; \mu_l, \Sigma_l)}$$

Averaging over all noise hypotheses $n_t^{(j)}$ according to their likelihood yields the estimate for x_t:

$$\hat{x}_t = \frac{\sum_{j=1}^{N} p(y_t|n_t^{(j)}) \cdot \hat{x}_t^{(j)}}{\sum_{j=1}^{N} p(y_t|n_t^{(j)})} = \sum_{j=1}^{N} \omega_t^{(j)} \cdot \hat{x}_t^{(j)}$$

The computational cost of the compensation is

$$\#particles \cdot \#gaussians \cdot \#(spectral\ bins)$$

which in praxis dominates the computational cost of the particle filter though its asymptotic complexity is the same.

2.3. Handling particle filter divergence

A well-known problem with tracking algorithms is deviation from the target trajectory, which sometimes cannot be recovered. We call this the '*divergence*' problem. Analogically, the PF for speech feature enhancement is said to diverge if the noise sequence is lost for an extended period of time which is usually accompanied by a continual and considerable misestimation of the noise. Substantial overestimations lead to severe problems with the likelihood computations since

$$\log(1 - e^{n_t^{(j)} - y_t})$$

in (2) cannot be computed if the magnitude of a noise hypothesis exceeds the corrupted speech spectrum. This is a consequence of considering the noise spectra — this time not in the log domain — to be additive, $y_t = x_t + n_t$. While Singh and Raj [7] have not addressed the problem, Haeb-Umbach and Schmalenstroeer [6] set the likelihood to 0. Furthermore, they report that this might lead to a severe decimation of the particle population up to its complete annihilation. We have also experienced the latter problem and handled it by repeating the first step of the PF(see section 2.1), thus by reinitializing the particles according to the noise distribution $p(n)$ if the overall likelihood

$$\sum_{j=1}^{N} p(y_t|n_t^{(j)})$$

got very small for a contiguous period of time (in our case 100 ms). The corresponding estimated clean speech spectra were replaced by their corrupted, non-filtered counterparts.

2.4. Improving noise robustness

Traditionally, PFs for the improvement of speech features are applied on the logarithmic mel power spectrum obtained by a *mel filterbank* on the power spectrum. The disadvantage of this approach — focusing on robust features — is the equal weighting of spectral peaks and valleys as it is well known that noise is mainly present in low energy regions. To overcome this drawback we estimate the spectrum by the warped and scaled *minimum variance distortionless response* (MVDR) [9] spectral envelope as it provides an accurate description only for spectral peaks. For the representation of valleys no information about the fine spectral structure is preserved, limiting the description more or less to the energy levels. Therefore, spectral envelopes are more robust to noise than their power spectrum counterparts. The MVDR is used instead of the widely known *linear prediction* (LP) as it has been shown that MVDR spectral estimation overcomes the problems in modeling voiced speech associated with LP spectral estimation techniques [10]. To provide a better approximation of the relevant aspects of the human auditory system, we have applied the well-known technique of pre-warping — a time-domain technique to estimate an all-pole model based on a warped frequency axis such as the mel scale — to the MVDR spectral estimate.

3. Coupling Particle Filters with Automatic Speech Recognition

Previous works used a general and static speech model (1) which systematically ignores the dynamic properties of speech. To overcome this deficiency we propose to use a time- and phoneme-

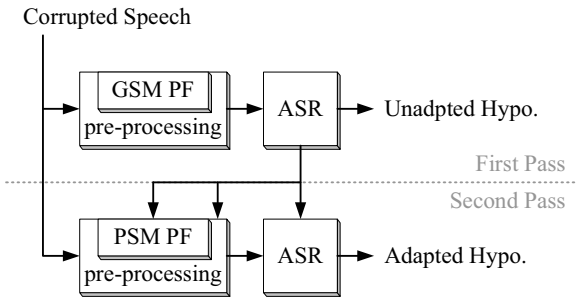

Figure 1: Flowchart of the coupling between the *particle filter* (PF) and *automatic speech recognition* (ASR) engine. GSM denotes *general speech model* and PSM denotes *phoneme-specific model*.

specific model

$$p_{phon(t)}(x) = \sum_{k=1}^{K} c_{k,phon(t)} \mathcal{N}(x; \mu_{k,phon(t)}, \Sigma_{k,phon(t)})$$

where $phon(t)$ denotes the phoneme spoken at time t. Since the phoneme is not known in advance, we use a 'two-pass' PF as depicted in Figure 1. In the first pass the PF with the general speech model is used to clean the speech spectra which are then processed with the speech recognition system to obtain a first phoneme sequence hypothesis (transcription). In the second (and following) pass(es), the hypothesis of the previous pass enables us to use the PF with the phoneme-specific model. This way the sophisticated acoustic and language models of the speech recognizer are incorporated into the particle filtering stage. Unfortunately, the phoneme-specific filter introduces two new problems:

1. Switching between phonemes causes a very sudden change of the particles' (noise hypotheses') likelihoods which can destabilize the PF.
2. By correcting all corrupted speech spectra toward the hypothesis from the previous pass we might tie ourselves too strongly to that hypothesis.

To overcome those problems we interpolate the phoneme-specific and the general model to form the *mixture model*

$$p_{mix(t)}(x) = \alpha \cdot p_{phon(t)}(x) + (1 - \alpha) \cdot p(x)$$

where α denotes the mixture weight. We have used an equal weighting between the general and phoneme-specific model throughout our experiments. A different value of α might lead to better results.

4. Speech Recognition Experiments

In order to evaluate the performance of the proposed PF enhancements under realistic conditions we have chosen approximately 45 minutes of lecture speech which present significant challenges to both modeling components used in *automatic speech recognition* (ASR), namely the language and the acoustic models. With respect to the former, the currently available lecture data primarily concentrates on technical topics with focus on speech and vision research. This is a very specialized task that contains many acronyms and therefore is quite mismatched to typical language models currently used in the ASR literature. Furthermore, large portions of the data contain spontaneous, disfluent, and interrupted speech, due to the interactive nature of seminars and the varying degree of the speakers' comfort with their topics. On the acoustic modeling side, and in addition to the latter difficulty, the seminar speakers exhibit moderate to heavy German or other European accents in their English speech. Part of this data has been used in the Rich Transcription 2005 Spring Meeting Recognition Evaluation [11].

To perform experiments with different *signal to noise ratios* (SNR)s we artificially added dynamic noise with a broad variety of sounds coming from a truck, slamming rubbish containers, distant voices, and shouts [12].

As a speech recognition engine we have used the *Janus Recognition Toolkit* (JRTk), which is developed and maintained by the Interactive Systems Laboratories at two sites: Universität Karlsruhe (TH), Germany and Carnegie Mellon University, USA. Relatively little supervised in domain data is available for acoustic modeling of the recordings. Therefore, we decided to train the acoustic model on the close talking channel of meeting corpora and merge it with the *Translanguage English Database* (TED) corpus [13] 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 [9]. Thereafter, linear discriminant analysis was used to reduce the utterance based cepstral mean normalized features plus 7 adjacent to a final feature number of 42. The acoustic model after merge and split training consisted of approximately 3,500 context dependent codebooks with up to 64 Gaussians with diagonal covariances each, summing up to a total of approximately 180,000 Gaussians. To train a 3-gram language model we have used corpora consisting of broadcast news, proceedings of conferences such as ICSLP, Eurospeech, ICASSP, ACL and ASRU and TED. The vocabulary contains approximately 23,000 words, the perplexity is around 125 with an out of vocabulary rate below 1.5%.

Table 1 shows *word errors rates* (WER)s for unadapted and adapted passes. In the second – adapted – pass, *maximum likelihood linear regression* (MLLR) [14] and constrained MLLR (feature space adaptation) [15] adaptation have been used on the hypotheses of the first – unadapted – pass. Vocal tract length normalization has not been used. The following discussion concentrates, if not stated otherwise, on the more relevant adapted results only.

For clean features the two different front-ends perform equally well. For decreasing SNRs the MVDR based features clearly outperform the Fourier based ones. The 'traditional' PF shows good improvements for the unadapted recognition pass which is reduced to marginal improvements on the adapted recognition pass. At 0 dB, the unfiltered MVDR based features can even improve accuracy over Fourier based ones that were cleaned using a 'traditional' PF. The combination of MVDR and 'traditional' PF can further improve the good result.

As most of the gain seen on the unadapted pass levels off on the adapted pass, we conclude that the adaptation of the speech recognition system compensates for most of the noise cleaned by the 'traditional' PF. The good result, a gain in accuracy of more than 5% relative, of the proposed phoneme-specific PF on the reference and the proposed mixture on the reference and hypotheses indicates, that the phoneme-specific PF is able to compensate for noises which can't be compensated for by the adaptation of the

			WER							
			clean speech		SNR 10 db		SNR 5 db		SNR 0 db	
front-end	filter type	PF adp. on	unadp.	adp.	unadp.	adp.	unadp.	adp.	unadp.	adp.
Fourier	none	-	31.7%	25.5%	42.6%	30.3%	48.7%	34.2%	62.7%	44.7%
warped MVDR	none	-	31.0%	25.4%	39.4%	29.2%	48.1%	33.8%	60.2%	42.4%
Fourier	GSM	-	-	-	41.0%	29.6%	46.2%	33.7%	60.1%	43.8%
warped MVDR	GSM	-	-	-	38.5%	28.4%	45.6%	33.5%	57.0%	42.1%
warped MVDR	PSM	reference	-	-	36.9%	28.3%	41.9%	30.0%	51.0%	36.7%
warped MVDR	PSM	hypotheses	-	-	-	28.5%	-	34.7%	-	43.0%
warped MVDR	MM	reference	-	-	37.1%	28.5%	43.7%	32.2%	53.9%	39.5%
warped MVDR	MM	hypotheses	-	-	-	28.4%	-	31.8%	-	40.3%

Table 1: *Word error rates* (WER)s for different front-ends, different or no *particle filter* (PF) and *signal to noise ratios* (SNR)s. The PF can either use the *general speech model* (GSM), the *phoneme-specific speech model* (PSM) or the *mixture model* (MM). PF adaptation of the PSM is either based on the hypothesis (unadapted recognition output) or the reference. The adapted speech recognizer pass has always been adapted with the output of the corresponding unadapted recognition pass.

speech recognition system. Note that the phoneme-specific PF failed in the case where no mixture model was used on the hypotheses of the ASR engine. This demonstrates the problem of 'model tying' as mentioned before.

Approximately 3 to 5 percent of the frames were lost because all particles had zero likelihood. The number of "dropouts" seamed to increase for a decrease in SNR and was 10 percent higher for features obtained by Fourier transformation than for the ones obtained by warped MVDR.

5. Conclusions

We have successfully demonstrated the combination of robust speech feature extraction in combination with statistical speech feature enhancement. Furthermore, we have coupled the independent treatment of particle filtering and speech recognition by using phoneme-specific models.

In the future we want to investigate different mixture weights for the phoneme-specific and general speech model, cluster similar phonemes and smooth the phoneme transitions.

6. Acknowledgment

The work presented here was partly funded by the *European Union* (EU) under the project CHIL (Grant number IST-506909).

7. References

[1] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," *IEE Proceedings on Radar and Signal Processing*, vol. 140, pp. 107–113, Sept. 1993.

[2] M. Isard and B. Blake, "Condensation – conditional density propagation for visual tracking," *Int. J. Computer Vision*, vol. 29, 1, pp. 5–28, 1998.

[3] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," *Proc. of ICASSP*, 2004.

[4] M. Fujimoto and S. Nakamura, "Particle filter based non-stationary noise tracking for robust speech feature enhancement," *Proc. of ICASSP*, 2005.

[5] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *ASSP*, vol. 27, pp. 113–120, Apr. 1979.

[6] R. Haeb-Umbach and J Schmalenstroeer, "A comparison of particle filtering variants for speech feature enhancement," *Proc. of Interspeech*, 2005.

[7] B. Raj and R. Singh, "Tracking noise via dynamical systems with a continuum of states," *Proc. of ICASSP*, 2003.

[8] P.J. Moreno, B. Raj, and R.M. Stern, "A vector taylor series approach for environment-independent speech recognition," *Proc. of ICASSP*, 1996.

[9] M. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[10] M.N. Murthi and B.D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 221–239, May 2000.

[11] NIST, "Rich transcription 2005 spring meeting recognition evaluation," *www.nist.gov/speech/tests/rt/rt2005/spring*.

[12] The Freesound Project, "garbage.coll.serv.ds70p.mp3," *freesound.iua.upf.edu/samplesViewSingle.php?id=6986*.

[13] Linguistic Data Consortium (LDC), "Translanguage english database," www.ldc.upenn.edu/Catalog/ LDC2002S04.html.

[14] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, pp. 171–185, 1995.

[15] M.J.F. Gales, "Semi-tied covariance matrices," *Proc. of ICASSP*, 1998.