INVENTORY-STYLE SPEECH ENHANCEMENT WITH UNCERTAINTY-OF-OBSERVATION TECHNIQUES

R. M. Nickel[1]*, R. F. Astudillo*[2]*, D. Kolossa*[3]*, S. Zeiler*[3]*, R. Martin*[3]

[1]Department of Electrical Engineering
Bucknell University, Lewisburg, PA 17837, USA
[2]INESC-ID Lisboa, Spoken Language Systems Lab
R. Alves Redol 9, 1000-029 Lisboa, Portugal
[3]Institut für Kommunikationsakustik, ID 2/231
Ruhr-Universität Bochum, 44780 Bochum, Germany
Primary Email: robert.nickel@bucknell.edu

ABSTRACT

We present a new method for inventory-style speech enhancement that significantly improves over earlier approaches [1]. Inventory-style enhancement attempts to resynthesize a clean speech signal from a noisy signal via corpus-based speech synthesis. The advantage of such an approach is that one is not bound to trade noise suppression against signal distortion in the same way that most traditional methods do. A significant improvement in perceptual quality is typically the result. Disadvantages of this new approach, however, include speaker dependency, increased processing delays, and the necessity of substantial system training. Earlier published methods relied on a-priori knowledge of the expected noise type during the training process [1]. In this paper we present a new method that exploits uncertainty-of-observation techniques to circumvent the need for noise specific training. Experimental results show that the new method is not only able to match, but outperform the earlier approaches in perceptual quality.

Index Terms— Inventory-Style Speech Enhancement, Uncertainty-of-Observation Techniques, Modified Imputation.

1. INTRODUCTION

Traditional speech enhancement methods based on filtering and/or spectral subtraction have reached a high level of sophistication. Yet, they are still far from matching a human's ability to separate speech from noise within an acoustic scene. Further technical improvements are possible, if we learn how to infuse more knowledge about the specific characteristics of speech into the enhancement process. *Inventory-style speech enhancement* provides a powerful avenue to do just that by tightly restricting the enhanced signal to segments of prerecorded speech from a targeted individual. Thereby, the procedure shares desirable properties with corpus-based synthesis in producing a high quality, naturally sounding output.

This advantage, however, comes at the price of a significantly increased complexity, higher storage requirements, speaker dependency, and substantial off-line training as implied in the method proposed by Xiao and Nickel in 2010 [1]. Specific techniques for the reduction of complexity and storage requirements for *their* method were discussed in [2]. In this paper we are presenting an approach that removes significant constraints in the off-line training of the system. While the original approach by Xiao and Nickel required noise specific training [1], we were able to devise a system that does not require any a priori knowledge about the expected noise during the training process. An expansion of the bandwidth of the proposed system from 8 kHz to 16 kHz was accomplished at the same time.

A key problem in inventory-style enhancement is the reliable recognition of the underlying phonetic class of a noisy speech segment. In standard approaches to speech recognition or phoneme classification, it is generally assumed that *preprocessing* delivers a good estimate of the clean speech signal. Preprocessed speech patterns are compared to the *clean* speech patterns on file to obtain an estimate of the true word or phoneme sequence. Under severely distorted conditions, however, using just a point estimate of the clean speech is generally not the optimum approach. Instead, it can be valuable to consider clean speech as an *unknown* process that can only be estimated with a residual, time-varying error variance. The ability to estimate this variance can lead to significant improvements in recognition performance, since the recognizer can then focus on more reliable segments and/or features via techniques such as variance compensation [3] or modified imputation [4]. In the specific case considered here, respective variances are estimated in the spectral domain and then propagated into a cepstral-feature domain with techniques described in [5] and [6].

2. PROPOSED ENHANCEMENT METHOD

In the proposed procedure we are dealing with discrete time signals that have been sampled at a rate of 16 kHz. We have access to a clean, i.e. undistorted, inventory of "example" recordings from the targeted individual. For notational simplicity we assume that all separately recorded inventory utterances are concatenated into one long signal stream $s[n]$. We use $x[n]$ to denote an observed noisy utterance from the targeted speaker. We consider $x[n] = z[n] + v[n]$ as the input to our enhancement system, in which $z[n]$ denotes the underlying clean speech signal and $v[n]$ denotes additive noise.

The proposed procedure employs signal segmentations with varying grades of "granularity". We use the notation $\mathbf{s}(n\,;L)$ to denote a vector of L successive samples of $s[n]$ from n to $n+L-1$, i.e. $\mathbf{s}(n\,;L) = [\ s[n]\ \ s[n+1]\ \ \ldots\ \ s[n+L-1]\ ]^{\mathrm{T}}$. The signal

[1]The first author performed the work while at the Institut für Kommunikationsakustik, Ruhr-Universität Bochum, Germany. The work was supported with a Marie-Curie International Incoming Fellowship through the European Union, Grant PIIF-GA-2009-253003-InventHI.

segmentation that is employed in the phonetic classification stage of our procedure (see Section 2.1) uses 20 msec frames with a 50 % overlap:

$$\mathbf{s}[i] = \mathbf{s}(160 \cdot i\,; 320) \tag{1}$$

The symbols $\mathbf{x}[i]$ and $\mathbf{z}[i]$ are defined analogously. The signal segmentation for the correlation-search stage of our procedure (see Section 2.3) operates on a finer grid. Here, we employ 10 msec frames with a 87.5 % overlap:

$$\mathbf{s}_k = \mathbf{s}(20 \cdot k\,; 160) \tag{2}$$

Again, symbols $\mathbf{x}_k$ and $\mathbf{z}_k$ are defined analogously. Enhancement is essentially performed by replacing incoming noisy frames $\mathbf{x}_k$ with clean inventory frames $\mathbf{s}_k$ as described in Section 2.3. A brief, conceptual description of our entire procedure, including the system training part, is provided in the following subsections.

2.1. Robust Feature Extraction

To increase the robustness in the unit selection process described in Section 2.3 a speech feature computation with uncertainty propagation was used. Our implementation followed the lines of [6], using a *Wiener filter* based spectral estimator to compute *mel-frequency cepstral coefficients* (MFCC) as features $\mathbf{c_s}[i]$ in synchronization with our signal segmentation from equation (1). For this particular implementation amplitude based MFCCs with cepstral mean subtraction where used to attain improved performance. In an initial step, conventional speech enhancement in the short-time Fourier transform (STFT) domain was applied as well. In this domain the most commonly applied model is the complex Gaussian distortion model, from which Wiener or Ephraim-Malah filters are derived. For this model, the Wiener estimator has the complex Gaussian posterior distribution $p(S|X) = \mathcal{N}_{\mathbb{C}}\left(S; S^W, \lambda\right)$ in which S^W is the Wiener estimate of a clean Fourier coefficient of $s[n]$ and λ is the residual mean square error. Conventional speech enhancement transforms the point estimate of the filter S^W into the feature domain. Uncertainty propagation transforms the whole associated posterior distribution instead, which leads to a corresponding feature posterior. It can be demonstrated that the posterior for the selected features is well approximated by the Gaussian distribution

$$p(\,\mathbf{c_s}[i] \,|\, \mathbf{x}[i]\,) \approx \mathcal{N}\left(\mathbf{c_s}[i]; \boldsymbol{\mu}_{\mathbf{cs}}[i], \boldsymbol{\Sigma}_{\mathbf{cs}}[i]\right) \tag{3}$$

were the feature means $\boldsymbol{\mu}_{\mathbf{cs}}[i]$ and covariance matrices $\boldsymbol{\Sigma}_{\mathbf{cs}}[i]$ can be obtained by applying [5, Eqs. 6.7, 6.9, 6.14, 6.16, 6.38, and 6.39]. Estimates of feature means $\hat{\boldsymbol{\mu}}_{\mathbf{cz}}[i]$ and covariances $\hat{\boldsymbol{\Sigma}}_{\mathbf{cz}}[i]$ for the clean segments $\mathbf{z}[i]$ that are underlying to our noisy input segments $\mathbf{x}[i]$ are computed from the $\mathbf{x}[i]$ analogously.

2.2. System Training and Inventory Design

The goal of the system training and inventory design stage is two fold: (1) we need to divide the inventory into collections $\mathbb{S}_q$ of phonetically similar segments $\mathbf{s}(n_p\,; L_p)$ with varying lengths L_p for $p = 1 \ldots P_q$, and (2) we need to arrive at a statistical description that tells us which set of collections $\mathbb{S}_q$ is most likely to contain an inventory subsection that best matches the underlying clean frame $\mathbf{z}_k$ of an incoming noisy frame $\mathbf{x}_k$.

The division of the inventory $s[n]$ into the collections $\mathbb{S}_q$ is performed in a step-by-step fashion. First, all silent sections contained in $s[n]$ are removed. The non-silent part of the inventory is then divided into sections that each belong to one of 40 phonetic classes. In this work, we employed the phonetic transcription that was provided with our experimental database (see Section 3). If a phonetic transcription is not available one may also use the unsupervised clustering method described in [1]. Both methods work equally well.

Short-time MFCC features means $\boldsymbol{\mu}_{\mathbf{cs}}[i]$ after Section 2.1 are computed for all segments of the inventory. A Gaussian mixture model (GMM) with 3 mixtures and diagonal covariance structure is trained with the corresponding $\boldsymbol{\mu}_{\mathbf{cs}}$ vectors for each phonetic class. The GMMs are then used to reclassify each $\boldsymbol{\mu}_{\mathbf{cs}}$. Many vectors remain in the same class, yet a substantial number is reassigned. The underlying *inventory sections* are reassigned accordingly as well. We then divide the $\boldsymbol{\mu}_{\mathbf{cs}}$ vectors of each phonetic class into three subclasses via a Euclidean k-means algorithm. Again, the underlying inventory sections are reassigned accordingly. The purpose of this subdivision is to provide flexibility in distinguishing coarticulation effects in the incoming speech. Lastly, we fit a Gaussian PDF model with a diagonal covariance structure to the $\boldsymbol{\mu}_{\mathbf{cs}}$ vectors of each of the resulting 120 phonetic[1] classes. As a result we obtain 120 Gaussian PDF models $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{cs}}; \boldsymbol{\mu}_{\mathbf{cs}q}, \boldsymbol{\Sigma}_{\mathbf{cs}q})$ and the associated signal segment collections $\mathbb{S}_q$ for $q = 1 \ldots 120$.

In a next step we consider the temporal development of the MFCC feature means $\boldsymbol{\mu}_{\mathbf{cs}}$ across the non-silent part of the inventory. Every vector $\boldsymbol{\mu}_{\mathbf{cs}}[i]$ is classified with the Gaussian PDF models $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{cs}}; \boldsymbol{\mu}_{\mathbf{cs}q}, \boldsymbol{\Sigma}_{\mathbf{cs}q})$ to belong to a specific class $q^*[i] \in [1, 120]$. All transitions from class $q^*[i]$ to class $q^*[i+1]$ are tallied and a 120×120-dimensional state transition probability matrix $\mathbf{P}$ is computed.

2.3. Inventory-Based Speech Enhancement

The enhancement process is performed in two separate streams $\hat{z}_F[n]$ and $\hat{z}_I[n]$, which are eventually merged to form the enhanced output signal $\hat{z}[n]$. Stream $\hat{z}_F[n]$ is computed from input frames $\mathbf{x}[i]$ according to a standard, filtering-based enhancement method. In our approach we use the popular *Log-MMSE* method by Ephraim and Malah in combination with the usual decision-directed approach for the estimation of the underlying a-priori SNR in the short-time DFT domain [7]. The resulting a-priori and a-posteriori SNR estimates are also used as an input to a log-likelihood style *voice activity detection* after Sohn and Kim [7, section 11.2] which provides a "voice active" or "voice not active" decision for each $\mathbf{x}[i]$. The decision of the VAD for each frame is fed through a 7-tap median filter to smooth out decision-flickering at the activity boundaries.

Those frames $\mathbf{x}[i]$ that are flagged as "voice active" are subjected to an inventory search process. In a first step, MFCC feature means $\hat{\boldsymbol{\mu}}_{\mathbf{cz}}[i]$ and the associated covariance estimates $\hat{\boldsymbol{\Sigma}}_{\mathbf{cz}}[i]$ are computed according to Section 2.1. A *noise adapted* feature mean estimate $\hat{\boldsymbol{\mu}}'_{\mathbf{cz}}[i]$ is produced for each of the $q = 1 \ldots 120$ phonetic class models according to the *modified imputation* approach:

$$\hat{\boldsymbol{\mu}}'_{\mathbf{cz}}[i] = (\hat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{cz}}[i] + \boldsymbol{\Sigma}^{-1}_{\mathbf{cs}q})^{-1}\,(\hat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{cz}}[i] \cdot \hat{\boldsymbol{\mu}}_{\mathbf{cz}}[i] + \boldsymbol{\Sigma}^{-1}_{\mathbf{cs}q} \cdot \boldsymbol{\mu}_{\mathbf{cs}q}) \tag{4}$$

With $\hat{\boldsymbol{\mu}}'_{\mathbf{cz}}[i]$ we can find likelihood values $\lambda_q[i]$ for each phonetic class via:

$$\lambda_q[i] = \mathcal{N}(\hat{\boldsymbol{\mu}}'_{\mathbf{cz}}[i]; \boldsymbol{\mu}_{\mathbf{cs}q}, \boldsymbol{\Sigma}_{\mathbf{cs}q}) \tag{5}$$

The likelihood values are normalized across all q to sum to one, so that they can be interpreted as observation probabilities of the given frame $\mathbf{x}[i]$ under the assumption of the phonetic class q. In combination with our state transition matrix $\mathbf{P}$ from Section 2.2 it is then possible via a *Viterbi algorithm* to find the most likely sequence $q^*[i]$ of phonetic class memberships for a given sequence of "voice active" segments $\mathbf{x}[i]$. The sequence $q^*[i]$ is expanded into not only the single best, but the three top most probable class memberships $q_1[i]$, $q_2[i]$, and $q_3[i]$ via a 'one-step' log-probability state estimator. The corresponding inventory collections $\mathbb{S}_{q_1}$, $\mathbb{S}_{q_2}$, and $\mathbb{S}_{q_3}$ are merged

[1] The resulting classes are, in fact, representing sub-phonetic units at this stage. For simplicity we will, nevertheless, refer to them as phonetic units.

to form the subset of all inventory sections $\mathbb{S}[i]$ that are considered possible representations for the clean underlying speech signal $\mathbf{z}[i]$ contained in $\mathbf{x}[i]$.

The search for the best segment in $\mathbb{S}[i]$ is performed with a matched filter approach [1]. For the search we are moving from our coarse grid, implied in our $\mathbf{x}[i]$ notation, to the fine grid, implied in our $\mathbf{x}_k$ notation from equation (2). The resulting best-fitting inventory frame $\mathbf{s}_k$ is first energy normalized to match the corresponding energy in stream $\hat{z}_F[n]$ and then cross faded with adjacent frames into the inventory stream $\hat{z}_I[n]$.

In a last step we are merging the two streams $\hat{z}_F[n]$ and $\hat{z}_I[n]$ into the enhanced output $\hat{z}[n]$. During "voice active" sections stream $\hat{z}_I[n]$ is switched on, and during "silent" sections stream $\hat{z}_F[n]$ is switched on.

3. EXPERIMENTAL RESULTS

We evaluated the performance of the proposed method in comparison to three reference techniques: (a) a baseline system similar to the one proposed by Xiao and Nickel in [1], (b) the *Log-MMSE* method proposed by Ephraim and Malah [7], and (c) the *Multiband Spectral Subtraction* (MBSS) method proposed by Kamath and Loizou [7]. Out of the wealth of possible choices for our references we chose the *Log-MMSE* and the *MBSS* approach because they were the most competitive from all of the reference techniques studied in [1].

The speech data that we used in our experiments was taken from the `CMU_ARCTIC` database from the Language Technologies Institute at Carnegie Mellon University[2]. It consists of recordings from seven English speakers with 1132 phonetically balanced utterances each. Most utterances are between one and four seconds long. The database includes full phonetic transcriptions of all utterances with (roughly) 40 elementary phonetic units per speaker.

The employed additive noise stems from the `NOISEX` database from the Institute for Perception-TNO, The Netherlands Speech Research Unit, RSRE, UK[3]. In our study we used three types: (1) *white* noise, (2) *buccaneer jet cockpit* noise (to represent a stationary, non-white noise type), and (3) *speech babble* (to represent a non-stationary noise type). The noise was added to the speech data at signal-to-noise ratios (SNR) of 5 dB, 10 dB, and 15 dB, under consideration of the respective active speech level after ITU-T P.56.

Enhanced speech was produced after the proposed inventory-based modified-imputation (**Inventory MI**) technique described in Section 2. The available speech data was split into two disjoint sets: (1) a training set, which served as our inventory and consisted of 1082 utterances per speaker, and (2) a testing set for performance evaluations, which consisted of 50 utterances per speaker. **Log-MMSE** and **MBSS** results were computed from the 50 testing utterances as described in [7]. As a baseline reference for inventory-style enhancement we used a method similar to the one proposed in [1]. It employs vector-quantization for phoneme classification and is therefore referred to as **Inventory VQ**. The differences between the Inventory VQ method used here and the one proposed in [1] were: (1) we are operating on 16 kHz data instead of 8 kHz data, (2) we are using 40, instead of 50, phonetic clusters, and (3) we are training our system with 10 dB white noise only. An evaluation at a different SNR value and with a different noise type, therefore, establishes a mismatch scenario between training and testing for the Inventory VQ case. The modifications were necessary to guarantee a fair comparison between the Inventory VQ method and the Inventory MI method, which does *not* require any noise specific training.

[2]The corpus is available at <http://festvox.org/cmu_arctic/>.

[3]The noise is available at <http://spib.rice.edu/spib/select_noise.html>.

TABLE I – PCRA SCORES UNDER VARIOUS NOISE CONDITIONS

PCRA in % (White Noise)	5dB	10dB	15dB
Inventory VQ (Xiao/Nickel)	74.45	77.83	**78.88**
Inventory MI (Proposed)	**81.10**	**79.20**	75.37
PCRA in % (Jet Cockpit Noise)	5dB	10dB	15dB
Inventory VQ (Xiao/Nickel)	69.40	73.44	**76.80**
Inventory MI (Proposed)	**78.86**	**77.27**	73.66
PCRA in % (Babble Noise)	5dB	10dB	15dB
Inventory VQ (Xiao/Nickel)	62.17	68.66	**72.51**
Inventory MI (Proposed)	**80.24**	**77.22**	71.33

The performance of all considered methods was evaluated with three objective quality measures: the *Phonetic Cluster Recognition Accuracy* (PCRA), the *Perceptual Evaluation of Speech Quality* (PESQ, [8]), and the *Short-Time Objective Intelligibility* (STOI, [9]).

The PCRA is an indirect measure of quality that allows us to compare inventory-based enhancement methods. It is correlated with both quality and intelligibility, especially at lower SNR values. PCRA scores report how often the estimated phonetic cluster index of a noisy frame was found to match the true phonetic cluster index of the underlying clean frame. For the proposed Inventory MI method, phonetic cluster recognition is performed *only* for segments for which the employed VAD, as described in Section 2, signals a "voice active" state. PCRA scores are therefore computed for such voice active sections only. To obtain a fair comparison we ensured that the computation of PCRA scores for the Inventory VQ method was based on the same sections.

Table I shows the resulting PCRA scores from our experiments under the three considered SNR levels and the three considered noise types. The score of the best performing algorithm, for each scenario respectively, is shown in bold-face letters. The proposed Inventory MI system outperforms the Inventory VQ method in all 5 dB and 10 dB cases across all considered noise types. Particularly remarkable is the 10 dB white noise case in which the Inventory MI method, which does *not* require any noise specific training, was able to eke out a slight PCRA gain over the Inventory VQ method, which was *specifically* trained for the 10 dB white noise case. The PCRA gain was especially dramatic for 5 dB babble noise. Here, the proposed method was able to achieve an absolute improvement of over 18 %-points in PCRA score. The proposed method was, unfortunately, not able to improve PCRA scores in the 15 dB SNR scenarios. At such high SNR values, however, the losses in PCRA did not translate to losses in either perceptual quality or intelligibility as measured by PESQ and STOI (see Figure 1 and Table II).

It is important to point out that one may not interpret table I as a study of PCRA scores vs. signal-to-noise ratio. A comparison across different SNRs is meaningless. The PCRA computation is a function of the underlying SNR-dependent VAD. The segments that are flagged by the VAD are *different* for different SNR scenarios and therefore PCRA scores can be compared fairly between algorithms but *not* fairly across different SNRs. The VAD bias also explains why the Inventory MI recognition rates seem to, paradoxically, improve with decreasing SNR values.

The main target of our study was to improve the resulting enhanced signals *perceptually* (i.e. subjectively). The PESQ score, an ITU recommendation after Rix *et al.* [8], is one of the few objective quality measures that correlate well with the subjective quality of speech. The resulting PESQ scores from our experiments under the three considered SNR levels and the three considered noise types are

TABLE II – PESQ MEASURES
UNDER VARIOUS NOISE CONDITIONS

PESQ (White Noise)	5dB	10dB	15dB
Noisy Signal	1.35	1.69	2.07
Inventory VQ (Xiao/Nickel)	1.95	2.42	2.71
Log-MMSE (Ephraim/Malah)	2.11	2.48	2.76
MBSS (Kamath/Loizou)	1.51	2.15	2.64
Inventory MI (Proposed)	**2.25**	**2.60**	**2.84**
PESQ (Jet Cockpit Noise)	5dB	10dB	15dB
Noisy Signal	1.32	1.63	2.00
Inventory VQ (Xiao/Nickel)	1.55	2.13	2.53
Log-MMSE (Ephraim/Malah)	1.79	2.29	2.62
MBSS (Kamath/Loizou)	1.42	1.94	2.49
Inventory MI (Proposed)	**1.82**	**2.37**	**2.71**
PESQ (Babble Noise)	5dB	10dB	15dB
Noisy Signal	1.66	1.97	2.32
Inventory VQ (Xiao/Nickel)	1.36	1.88	2.33
Log-MMSE (Ephraim/Malah)	1.83	2.22	2.57
MBSS (Kamath/Loizou)	1.85	2.28	2.66
Inventory MI (Proposed)	**1.96**	**2.36**	**2.68**

shown in Table II. Again, the score of the best performing algorithm for each considered scenario is shown in bold-face letters.

It is readily seen that the proposed Inventory MI method outperforms all of the reference methods in all of the considered noise scenarios. The improvements over the Inventory VQ method were among the largest. The relatively poor performance of the Inventory VQ methods is primarily a consequence of: (1) the inherent mismatch between training and testing[4] and (2) the expansion of the considered signal bandwidth from 8 kHz to 16 kHz.

The improvements over the Log-MMSE and the MBSS method are less dramatic but still significant. It is known that the Log-MMSE method performs generally well in stationary noise, such as jet cockpit noise, whereas the MBSS method thrives in non-stationary noise, such as speech babble. Yet, neither of the two reference methods is able to obtain top scores in either scenario. The proposed technique outperforms both methods for both noise types. This result is expected since the proposed method has access to speech specific information that is not accessible to either the Log-MMSE or the MBSS method.

The maintenance of a high level of intelligibility is a concern that applies especially to enhancement methods that employ an automatic recognition of phonetic content in low SNR scenarios. An objective measure that is specifically designed to asses speech intelligibility is provided by the recently published STOI measure after Taal *et al.* [9]. The STOI measure is normalized between zero and one. Scores close to one indicate perfect intelligibility.

Figure 1 shows the resulting STOI scores from our experiments under the jet cockpit noise scenario. The STOI results for the other two noise scenario were similar, but had to be omitted here due to space limitations. The goal for our proposed method was the improvement of the perceptual quality of the enhanced signals without a noticeable reduction in intelligibility. Figure 1 shows that the STOI scores of the proposed method are never significantly below the respectively best STOI scores from all of the reference methods. This result was also confirmed by a few informal listening tests with expert listeners who rated the intelligibility of the enhanced signals as virtually undiminished.

[4]Except for the 10 dB white noise case.

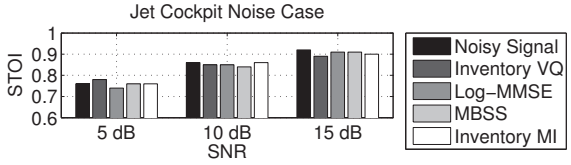

Figure 1. Average STOI scores for the jet cockpit noise case.

4. CONCLUSION

We presented a substantial revision of a previously published method for inventory-style speech enhancement [1]. The newly proposed method employs a phoneme recognition front-end that incorporates uncertainty-of-observation techniques into the enhancement process. With the new front-end it becomes possible to substantially relax the training requirements for the system. The training of the earlier system still depended on an a-priori knowledge of the expected noise type. The new approach, however, is no longer bound by that requirement. Experiments show that the new method is not only able to match, but outperform the earlier approach in *objectively* measured perceptual quality, while retaining the *objectively* measured intelligibility. Informal listening tests with expert listeners also confirmed these results. Sound samples will be provided at the conference.

5. REFERENCES

[1] X. Xiao and R. M. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 18, no. 6, pp. 1243–1257, Aug. 2010.

[2] R. M. Nickel and R. Martin, "Memory and complexity reduction for inventory-style speech enhancement systems," *Proc. of EUSIPCO, Barcelona, Spain*, pp. 196–200, Sept. 2011.

[3] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Sp. & Aud. Proc.*, vol. 13, no. 3, pp. 412–421, May 2005.

[4] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2005, pp. 82–85.

[5] R. F. Astudillo and R. Orglmeister, "A MMSE estimator in mel-cepstral domain for robust large vocabulary automatic speech recognition using uncertainty propagation," in *Proc. Interspeech*, 2010, pp. 713–716.

[6] Ramon Fernandez Astudillo, *Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition*, Ph.D. thesis, Technical University Berlin, 2010.

[7] P. C. Loizou, *Speech Enhancement – Theory and Practice*, CRC Taylor and Francis, 2007.

[8] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *Proceedings of ICASSP*, vol. 2, pp. 749–752, 2001.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.