Inventory-Based Audio-Visual Speech Enhancement

Dorothea Kolossa¹, Robert Nickel², Steffen Zeiler³, Rainer Martin¹

¹Institut für Kommunikationsakustik, ID 2/231
Ruhr-Universität Bochum, 44780 Bochum, Germany
²Department of Electrical Engineering
Bucknell University, Lewisburg, PA 17837, USA
dorothea.kolossa@rub.de, robert.nickel@bucknell.edu,
steffen.zeiler@rub.de, rainer.martin@rub.de

Abstract

In this paper we propose to combine audio-visual speech recognition with inventory-based speech synthesis for speech enhancement. Unlike traditional filtering-based speech enhancement, inventory-based speech synthesis avoids the usual trade-off between noise reduction and consequential speech distortion. For this purpose, the processed speech signal is composed from a given speech inventory which contains snippets of speech from a targeted speaker. However, the combination of speech recognition and synthesis is susceptible to noise as recognition errors can lead to a suboptimal selection of speech segments. The search for fitting clean speech segments can be significantly improved when audio-visual information is utilized by means of a coupled HMM recognizer and an uncertainty decoding framework. First results using this novel system are reported in terms of several instrumental measures for three types of noise.

Index Terms: audio-visual speech enhancement, speech synthesis, unit selection, missing data techniques

1. Introduction

Single-channel noise reduction algorithms are used in many applications to increase listening comfort or automatic speech recognition quality [1]. However, the disadvantage of traditional filtering-based methods is that they almost always trade the resulting level of residual noise against a certain level of speech distortion and, thereby, almost invariably reduce intelligibility [2].

A promising alternative to the traditional filtering and spectral subtraction approaches is provided by inventory-style, i.e. corpus-based, enhancement schemes. These methods employ an analysis/resynthesis paradigm in which the distorted speech signals are not just filtered, but, instead, subjected to a resynthesis procedure that creates an entirely new “clean” signal from a pre-recorded corpus of speech waveforms. The selected waveforms are chosen to match the estimated underlying speech content of the observed noisy signal as closely as possible. Technically feasible implementations of such an inventory-based speech processing (IBSP) scheme were published by Xiao and Nickel [3] and by Ming et al. [4]. A significant refinement of the original method [3] was proposed by Nickel et al. in 2012 [5]. The analysis/resynthesis paradigm has - at least in theory - the potential to produce clean speech signals that are free of artifacts and therefore highly intelligible.

One of the limitations of the original approach, however, lies in its reliance on the distorted audio signal. This problem can be ameliorated to a good extent by using missing data techniques [5]. Yet, under extremely difficult conditions, additional information may be needed for high-quality results. In this paper, we present an additional approach for filling in missing information: an audio-visual speech recognition system is used, which fills in gaps in audio information via the use of additional video features. It can compensate through “lip-reading” what is missing in the audio signal alone. The system is based on coupled HMM decoding with the added benefit of missing data techniques. It is described in Section 2.

The audio-visual recognition result, a frame-by-frame estimate of the phonetic state, is used for inventory-based speech processing in the next step. IBSP replaces the acoustic signal in each frame by the best fitting counterpart from the clean inventory of the respective audio state. This process is discussed in more detail in Section 3. It leads to a speech estimate that combines audio, video, and automatic speech recognition (ASR) grammar information to obtain a signal that is completely free of additive noise, as it is based solely on an inventory of clean speech. Due to its use of missing data techniques, the system does not require noisy training or adaptation data to be of great advantage in different types of noise environments, stationary and nonstationary alike.

The evaluation of the proposed method, including the generation of noisy test data, an introduction of the applied quality measures, and a summary of the resulting performance analysis is given in Section 4. Conclusions are drawn in Section 5.
2. Audio-Visual Speech Recognition

For recognition of multimodal, or specifically audio-visual, speech recordings, it is necessary to extend the standard HMM framework to allow for varying time-delays between modalities. This is due to speakers visibly bringing the articulators into position, before actually producing audible utterances, so that video information tends to precede audio data by up to 120ms [6].

2.1. Coupled HMMs

Coupled HMMs are a special case of dynamic Bayesian networks [7]. They describe the temporal evolution of the audio and visual speech components, making allowances for their natural asynchrony. Unlike the product HMM, which uses frame synchronous concatenations of audio and video observation vectors, the coupled HMM is not influenced by a certain amount of asynchrony between both modalities, and it enforces synchrony only at HMM model boundaries (see Fig. 1). In multi-stream HMMs, on the other hand, the coupling would be too loose to infer the audio state sequence from a common audio-visual state sequence, which is a requirement here for the inference of phoneme labels. Coupled HMMs are also economical in the number of parameters. Although the number of states is proportional to the Cartesian product of the marginal audio and video model states, the number of output density functions grows only linearly and is \(N + M\) for a model with \(N\) audio and \(M\) video states. Each word in the grammar is modeled by a CHMM with \(N = 3P\) audio and \(M = P\) video states, with \(P\) as the number of phonemes of the word. A stream weight \(\gamma\) is used to calculate the observation probability for the coupled state \(q\) from the audio observation \(o_a\) and video observation \(o_v\), given the audio state \(q_a\) and the video state \(q_v\), according to

\[
p(o|q) = p(o_a|q_a)\gamma \cdot p(o_v|q_v)^{(1-\gamma)}.
\]  

In all experiments, \(\gamma\) was set to 0.5, which gives an equal weight to audio and video streams. As audio observations \(o_a\), mel frequency cepstral coefficients with first and second derivatives are computed as in [8]. The video features \(o_v\) are 64-dimensional DCT coefficients of the mouth region, which is detected by the Viola-Jones algorithm [9].

2.2. Inventory Construction

To construct an inventory of speech signals associated with any given audio HMM state, it is necessary to divide and assign all training data to appropriate HMM states \(q_a\). If we assume that our data (for both training and testing) is sampled at \(f_s = 8\) kHz then we can write a 10 msec long segment of successive samples from our training data \(s[n]\) as

\[
s[n] = [ s[n-40] \, s[n-39] \, \ldots \, s[n+39]]^T.
\]  

Since the training data is clean data, the assignment of a unique state \(q_a\) to each frame \(s[n]\) can be performed using audio-based information only. Specifically, we carried out a forced alignment of the training data \(s[n]\) with the correct sequence of acoustic HMMs, where the sequence was determined from the data annotation. This forced alignment, or *Viterbi alignment*, is the same as used in standard HMM initializations, and it provides the maximum likelihood assignment of all feature vectors in the training set to the best-fitting audio HMM state \(q_a\). As a result we obtain a unique *frame-to-state* mapping \(q_a = ftsmap(n)\) of the entire set of training utterances. All audio frames with the same label \(q_a\) are then grouped into the inventory \(S_{q_a}\) of the audio-HMM state that they are associated to, i.e.

\[
S_{q_a} = \{ s[n] \mid q_a = ftsmap(n) \}.
\]  

In this manner, we obtain an inventory of speech data that can serve as a waveform repertoire for each audio HMM state.

2.3. Coupled HMM Decoding for Audio-Visual Speech State Labeling

In order to process a noisy recording, the coupled HMM is used in a two-stage process. The first stage, audio-visual decoding, recognizes the sequence of words that were spoken in the recorded sentence. In the second stage, the forward-backward algorithm is used for a precise frame-state alignment, which is needed for the inventory-based processing described in Section 3.

2.3.1. Audio-Visual Decoding

For audio-visual speech recognition, coupled HMMs are used for each word in the language model, and are composed to form a search network, which is traversed in the standard token passing framework [10] by means of the JASPER audio-visual speech recognition system [11].
One special feature of JASPER is its support for uncertainty-of-observation techniques, which enhance robustness by considering the speech feature vectors not as representatives of the clean speech itself, but rather, as clean speech estimates, which are only given with a time-varying variance \( \Sigma_j \), where \( j \) indicates the frame index of the recognition features. By estimating \( \Sigma_j \) on a frame-by-frame basis, we can then focus most on the more reliable segments of audio data, and, during unreliable phases with large \( \Sigma_j \), decrease the influence of the distorted audio input on the recognition output.

For this purpose, we modify the computation of HMM state observation probabilities from standard Gaussian mixture models

\[
p(o_a | q_a) = \sum_{m=1}^{M} w_{q_a,m} \cdot N(o_a, \mu_{q_a,m}, \Sigma_{q_a,m}) \tag{4}
\]

with \( \mu_{q_a,m} \) and \( \Sigma_{q_a,m} \) as the mean and the covariance matrix of mixture \( m \) in state \( q_a \) and \( w_{q_a,m} \) as the associated mixture weight. Instead of using (4), the audio observation probabilities are computed by modified imputation [11] according to

\[
p(o_a | q_a) = \sum_{m=1}^{M} w_{q_a,m} \cdot N(\hat{o}_a, \mu_{q_a,m}, \Sigma_{q_a,m}). \tag{5}
\]

Here, \( \hat{o}_a \) is a model-based frame- and state-dependent estimate of the clean speech feature vector, given by

\[
\hat{o}_a = (\Sigma_{q_a,m} + \Sigma_j)^{-1} (\Sigma_j \mu_{q_a,m} + \Sigma_{q_a,m} o_a).
\]

The cepstrum-domain uncertainties \( \Sigma_j \) are estimated from the variances of a Wiener filter in the STFT domain [12], which are propagated to the MFCC domain as described in [8].

2.3.2. Inventory State Labeling

Based on the audio-visual recognition output - the word sequence and the alignment between words and audio feature vectors - the forward-backward algorithm is used to obtain a precise state-frame alignment. In this way, the best-fitting audio HMM state \( q_a \) can be determined for each frame, which then forms the basis for the actual inventory-based speech enhancement.

3. Inventory-Based Speech Processing

Inventory-based speech processing allows us to convert the implied phonetic information that is extracted via the CHMM from an incoming noisy signal \( x[n] \) back into an enhanced signal output waveform \( y[n] \). To that end, we are segmenting our noisy signal \( x[n] \) into 10 msec long processing frames with a 50% overlap, i.e.

\[
x_i = [ x[40i - 40] \ x[40i - 39] \ldots \ x[40i + 39] ]^T.
\]

Similarly to Section 2.2, we obtain an implied mapping \( q_a = \text{ftsmap}(40i) \) from our CHMM which assigns a unique state \( q_a \) to each overlapping frame \( x_i \). In this case, however, the CHMM estimates the states \( q_a \) by combining the information extracted from the noisy audio track \( x[n] \) with the associated visual information as described in Section 2.3.

A reconstruction waveform \( y'_i \) is found for each noisy input frame \( x_i \) via a normalized correlation procedure\footnote{Matched filter approach} across all waveforms of the associated inventory \( \mathbb{S}_{q_a} \):

\[
y'_i = \underset{s \in \mathbb{S}_{q_a}}{\text{arg max}} \left| \frac{x_i^T \cdot s}{\|x_i\| \cdot \|s\|} \right| \text{ for } q_a = \text{ftsmap}(40i).
\]

The correlation search for frames \( x_i \) that are near the transition to a different state is not only conducted across the inventory \( \mathbb{S}_{q_a} \) of the current state but also across the inventory of the followup state to smooth out state estimation errors at the transition boundaries.

The targeted energy of the desired output frame \( y \) is estimated by assuming that speech and noise are uncorrelated, i.e.

\[
y = \text{sign}(x_i^T \cdot s) \cdot \sqrt{\text{ppos}(\|x_i\|^2 - \sigma_n^2)} / \|y'\| \cdot y'. \tag{6}
\]

The power of the noise \( \sigma_n^2 \) can be obtained from frames that are flagged as silent by the CHMM. The positive part only-function \( \text{ppos}(q) = \frac{1}{2} (|q| + q) \) ensures that the energy estimate is not negative. The resynthesis of the targeted enhanced output signal \( y[n] \) from the frames \( y \) is accomplished with the interlacing cross-fading scheme that is comprehensively described in [3]. As a post-processing step, all segments which are labeled as silence by the CHMM system are damped by a factor of \( \frac{1}{100} \).

4. Experiments and Results

4.1. Experimental Setup

Data from the GRID corpus [13] is used with artificially added noise. We used three types of noise: (1) white noise, (2) buccaneer jet cockpit noise, and (3) speech babble. The noise signals were taken from the NOISEX database from the Institute for Perception-TNO, The Netherlands Speech Research Unit, RSRE, UK\footnote{The noise is available at <http://spib.rice.edu/spib/select_noise.html>}. We chose buccaneer jet cockpit noise as an example for a stationary, non-white noise type and speech babble as an example for a non-stationary noise type. The noise was added to the speech data at a signal-to-noise ratio (SNR) of 10 dB after ITU-T P.56 [14].

4.2. Evaluation

For the evaluation, we have used a range of objective quality measures that cover spectral and cepstral features as well as an intelligibility measure: the cepstrum distance was computed according to \[15\], the segmental...
SNR, defined on page 45 of [16], was used in the implementation provided by Loizou [1], and the short-term objective intelligibility measure (STOI) described in [17] was also determined. We compared the widely used log-spectral amplitude MMSE estimator [1] with our results based on audio-only recognition (ASR-IBSP) and audio-visual recognition (AVSR-IBSP). Results for all methods are shown in Tables 1 to 3, with best results indicated in bold print.

### Table 1: Cepstral Distance, ideally 0.

<table>
<thead>
<tr>
<th></th>
<th>unprocessed</th>
<th>log-MMSE</th>
<th>ASR-IBSP</th>
<th>AVSR-IBSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>0</td>
<td>1.31</td>
<td>3.80</td>
<td>3.80</td>
</tr>
<tr>
<td>white</td>
<td>7.53</td>
<td>6.96</td>
<td>5.11</td>
<td>4.88</td>
</tr>
<tr>
<td>jet</td>
<td>5.91</td>
<td>5.69</td>
<td>4.42</td>
<td>4.31</td>
</tr>
<tr>
<td>babble</td>
<td>5.05</td>
<td>5.15</td>
<td>5.12</td>
<td>4.77</td>
</tr>
</tbody>
</table>

### Table 2: Segmental SNR, ideally ∞ but clipped to 35 in the computation to allow averaging.

<table>
<thead>
<tr>
<th></th>
<th>unprocessed</th>
<th>log-MMSE</th>
<th>ASR-IBSP</th>
<th>AVSR-IBSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>35.0</td>
<td>14.9</td>
<td>7.3</td>
<td>7.1</td>
</tr>
<tr>
<td>white</td>
<td>-0.9</td>
<td>3.1</td>
<td>4.4</td>
<td>4.7</td>
</tr>
<tr>
<td>jet</td>
<td>-0.9</td>
<td>2.1</td>
<td>4.1</td>
<td>4.2</td>
</tr>
<tr>
<td>babble</td>
<td>-0.6</td>
<td>1.7</td>
<td>3.3</td>
<td>3.5</td>
</tr>
</tbody>
</table>

### Table 3: STOI, ideally 1.0.

<table>
<thead>
<tr>
<th></th>
<th>unprocessed</th>
<th>log-MMSE</th>
<th>ASR-IBSP</th>
<th>AVSR-IBSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>1.00</td>
<td>0.80</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>white</td>
<td>0.61</td>
<td>0.60</td>
<td>0.65</td>
<td>0.67</td>
</tr>
<tr>
<td>jet</td>
<td>0.61</td>
<td>0.60</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>babble</td>
<td>0.63</td>
<td>0.61</td>
<td>0.63</td>
<td>0.64</td>
</tr>
</tbody>
</table>

As it can be seen, some distortions still occur for clean speech. In noisy conditions, however, the suggested approach clearly outperforms the traditional log-MMSE estimator and the audio-only recognition, in terms of all considered quality measures.

### 5. Conclusions

A cross-modal speech-enhancement system has been presented, which offers a good trade-off between noise reduction and speech distortion in noisy environments. In contrast to many standard approaches, it offers not only an increased SNR and a reduced cepstral distortion, but also shows the potential to improve intelligibility, as indicated by our STOI measurements. Assuming good visual conditions, the performance at high acoustic SNR conditions is bounded by the distortions introduced by the inventory-based synthesis model and at very low SNR conditions by the visual-only performance of the AVSR system. As we like to apply this system in noisy conditions, the best possible “clean” performance was not the main concern of this work. The extension of this approach to very low SNR conditions, however, hinges on further improvements of the AVSR system, especially on the visual component. This is a challenging task which will be explored in future works.

### 6. Acknowledgments

This work has been supported by the Ministry of Economic Affairs and Energy of the State of North Rhine- Westphalia, Grant IV.5-43-02/2-005-WBFO-009, and the European Union, Grant PIIF-GA-2009-253003-InventHI.

### 7. References


INTERSPEECH 2012 590