INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Introducing the Turbo-Twin-HMM for Audio-Visual Speech Enhancement

Steffen Zeiler[1], *Hendrik Meutzner*[1], *Ahmed Hussen Abdelaziz*[2], *Dorothea Kolossa*[1]

[1]Cognitive Signal Processing Group, Ruhr-University Bochum, Germany
[2]International Computer Science Institute (ICSI), Berkeley
{ steffen.zeiler, hendrik.meutzner, ahmed.hussenabdelaziz, dorothea.kolossa }@rub.de

Abstract

Models for automatic speech recognition (ASR) hold detailed information about spectral and spectro-temporal characteristics of clean speech signals. Using these models for speech enhancement is desirable and has been the target of past research efforts. In such model-based speech enhancement systems, a powerful ASR is imperative. To increase the recognition rates especially in low-SNR conditions, we suggest the use of the additional visual modality, which is mostly unaffected by degradations in the acoustic channel. An optimal integration of acoustic and visual information is achievable by joint inference in both modalities within the turbo-decoding framework. Thus combining turbo-decoding with Twin-HMMs for speech enhancement, notable improvements can be achieved, not only in terms of instrumental estimates of speech quality, but also in actual speech intelligibility. This is verified through listening tests, which show that in highly challenging noise conditions, average human recognition accuracy can be improved from 64% without signal processing to 80% when using the presented architecture.
Index Terms: audio-visual speech enhancement, turbo decoding, twin-HMM

1. Introduction

The Turbo-Twin-HMM is a new model for audio-visual speech enhancement. It is a crossover between the *Twin-HMM* [1] for speech synthesis and the turbo principle [2] for audio-visual speech decoding. Audio-visual speech enhancement uses the additional visual modality to recover speech information from a noise-corrupted audio channel. A good audio recognizer in combination with a decent lip-reading system clearly exceeds the performance of audio-only methods [3].

The *Twin-HMM* for audio-visual speech enhancement [4] uses the best audio state sequence through a coupled HMM [5] to find an estimate of the clean speech signal. If the decoder accidentally selects an incorrect path, the synthesized signal will resemble a noise free but most certainly wrong-sounding phoneme. Instead of the coupled HMM decoder in the following we use the turbo decoder to calculate the required state posteriors. The turbo principle for audio-visual speech decoding uses all available audio and video sequence information over multiple iterations for the recognition and is therefore advantageous compared to the left-right Viterbi style decoding of the coupled HMM. To alleviate the problem of hard decisions in the decoder, we propose a framewise mixture of all HMM output models, weighted by their state posteriors. The idea is that a wrong decision by the decoder will be partially compensated by the outputs of other models.

This work was partially supported by the EU FET grant TWO!EARS (ICT-618075).

Although this paper describes the pure synthesis of noise-free audio signals from a parametric audio-visual speech model, we are not limited to this application. If we treat acoustic observations as random variables and combine them with expectations and uncertainties [6] from the parametric model we can optimally combine the input and the modeled spectrum in an MMSE estimate [1].

In the following, we will give an overview of existing work on audio-visual speech enhancement in Section 2. Sections 3 and 4 review the *Twin-HMM* for speech enhancement and the turbo principle for audio-visual speech recognition. In Section 5 we introduce the new Turbo-Twin-HMM architecture for speech enhancement. Section 6 describes our experiments with the GRID database. In Sections 7 and 8 we give results for various instrumental speech quality measures and for human listening tests. The intelligibility improvements of the new Turbo-Twin-HMM are discussed in Section 9 before we conclude in Section 10.

2. Audio-Visual Speech Enhancement

Two basic difficulties arise in the application of ASR knowledge to speech enhancement. On the one hand, ASR models of speech typically work in feature domains that are highly discriminative for phonetic classes, and that therefore attempt to minimize the influence of the speaker traits, room characteristics, and any prosody. This makes ASR features such as MFCCs insufficiently descriptive of the speaker identity and speaking style for the purpose of speech enhancement.

On the other hand, ASR performance degrades rather quickly, once the acoustic environment is degraded by noise or reverberation, which implies that any ASR-based speech enhancement is also bound to deteriorate in noisy conditions.

These two issues have both been addressed in previous works. To compensate for the fact that ASR models do not have sufficient speaker identity and prosodic cues, mainly two approaches have been attempted: inventory-based [7, 8, 9] and Twin-HMM-based speech enhancement [4]. Regarding the performance in noisy or otherwise distorted conditions, a range of different approaches exists. The use of audio-visual data has been investigated in [1, 4]. Greater ASR robustness can also be achieved using observation uncertainties [10, 11] or through parallel model combination [12].

In the following, we will present a new approach that, like [4], is based on audio-visual recognition, using the idea of a *Twin-HMM*. However, in contrast to earlier work, the audio-visual decoder will now be based on the turbo principle instead of coupled HMMs, which has shown superior performance in, e.g., [13, 14] and we will investigate this new model architecture — the *Turbo-Twin HMM* — in terms of three instrumental measures, the PESQ, STOI and segmental SNR. In addition, we

http://dx.doi.org/10.21437/Interspeech.2016-350

will show listening test results that indicate large improvements in speech intelligibility for negative SNRs, which imply that our introduced model is capable of notably enhancing the actual intelligibility of noisy speech.

3. Twin-HMM

The *Twin-HMM* assigns two different output density functions (ODFs) to each HMM state. The first output density function corresponds to the usual ODF insofar as it is used to find the most likely state sequence within a recognition architecture, and is hence referred to as *REC*, because it models the system's recognition features.

The second set of features, the *synthesis features*, or *SYN* features, models the clean amplitude spectrum of speech and is hence useful for speech enhancement or synthesis. In the following, this second set of distributions will be used to estimate the clean amplitude spectrum of speech based on a recognized state sequence obtained from turbo decoding.

With the *Twin-HMM* it is hence possible to use one set of features and ODFs, e.g. MFCCs based on multicondition training, that are optimized for maximum phonetic discriminance and best recognition results, and another completely different set of features and ODFs that are most suitable to estimate the clean speech amplitude spectrum. In this way, the *Twin-HMM* provides a statistical model that is suitably structured to describe the co-evolution of two streams of data – one for recognition and the other for speech processing.

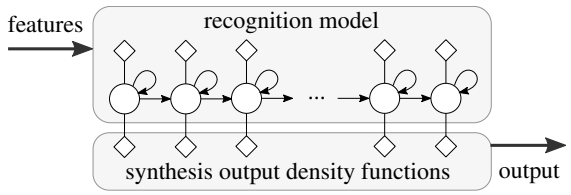

Figure 1: Concept of the *Twin-HMM* for model-based speech enhancement. A conventional HMM in the upper part is augmented by an additional set of ODFs for speech synthesis or reconstruction.

4. Turbo Decoding

Turbo decoding (TD) has been developed for convolutional error correction and channel decoding in digital transmission systems [15, 16]. Recently, TD was introduced into the field of ASR in order to perform multi-modal recognition [2, 13, 14].

TD is based on the iterative exchange of information, deduced from state posteriors, between different decoders. This extra information, g_a and g_v in Fig. 2, is used like a prior to modify the observation likelihoods b_a and b_v for the calculation of state posteriors in the forward-backward algorithm (FBA). The modified audio and video likelihoods for the respective observations o_a and o_v are

$$\tilde{b}_a(o_a|q_a) = b_a(o_a|q_a) \cdot g_a(q_a)^{\lambda_T \lambda_P}, \quad (1)$$

$$\tilde{b}_v(o_v|q_v) = b_v(o_v|q_v) \cdot g_v(q_v)^{(1-\lambda_T)\lambda_P}, \quad (2)$$

in which the constant λ_P balances the likelihoods and prior probabilities and λ_T is used like a conventional audio stream weight. For an extended discussion of adaptive stream weighting for coupled HMMs and TD, see [17, 18].

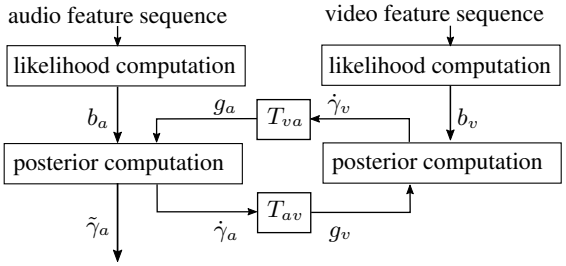

Figure 2: Turbo decoding for AVSR. The left column is a conventional audio-only ASR system. A decoder for the second modality (video) is added and the extrinsic probabilities $\dot{\gamma}_a$ and $\dot{\gamma}_v$ are exchanged iteratively between the two decoders.

From the FBA, we obtain new state posteriors $\tilde{\gamma}$, which subsume the likelihood, the prior probability and the extrinsic probability [13]. To find the extrinsic probability $\dot{\gamma}(q(t))$ for state q and frame t, we have to remove all excess information via

$$\dot{\gamma}\big(q(t)\big) \propto \frac{\tilde{\gamma}\big(q(t)\big)}{b\big(o(t)|q(t)\big) \cdot g\big(q(t)\big)}. \quad (3)$$

The extrinsic probabilities $\dot{\gamma}$ are mapped to the other decoder's state space by a linear transformation T_{va} or T_{av} respectively.

$$g_a = T_{va}\, \dot{\gamma}_v \qquad \text{audio} \leftarrow \text{video} \quad (4)$$

$$g_v = T_{av}\, \dot{\gamma}_a \qquad \text{video} \leftarrow \text{audio} \quad (5)$$

In our experiments the process of FBA followed by the deduction of extrinsic probabilities and their transfer to the corresponding other state space is iterated 4 times. In the first TD iteration, a flat prior $g_a(q_a) = 1, \ \forall q_a$ is used for the audio states. After the final iteration TD finishes with the audio posteriors $\tilde{\gamma}_a$.

5. Turbo-Twin-HMM

The Turbo-Twin-HMM joins the turbo principle for multimodal speech decoding and the *Twin-HMM* for speech reconstruction. Joint inference in the audio and video sequence of an utterance is done via turbo decoding. After a few iterations we obtain audio state posteriors as a prerequisite for the clean speech estimation. To find an estimate of the clean speech, we have attached an additional set of SYN output density functions to all states of the audio model according to the *Twin-HMM* principle.

Here we compare two ways of synthesis, all-path (AP) and best-path (BP) synthesis. For AP synthesis, we calculate the minimum-mean square error estimate of the clean speech amplitude spectrum $\hat{x}_{\text{AP}}(t)$ as

$$\mathrm{E}\big(x(t)|o(t)\big) = \sum_{i=1}^{N} p\big(q(t)=i|o(t)\big)\, \mathrm{E}\big(x(t)|q(t)=i\big). \quad (6)$$

This corresponds to a sum of synthesis ODF means μ_i over all states i weighted by the corresponding audio state posterior $\tilde{\gamma}_a(i,t)$ (see Fig. 2) for frame index t,

$$\hat{x}_{\text{AP}}(t) = \sum_{i=1}^{N} \tilde{\gamma}_a(i,t)\, \mu_i. \quad (7)$$

For BP synthesis, instead of a weighted sum, we use the most probable state $i^*(t)$ for each frame t, to compute the clean speech signal estimate

$$\hat{x}_{\text{BP}}(t) = \mu_{i^*(t)}. \quad (8)$$

Table 1: Instrumental measures for the noisy data, the audio-only log-MMSE speech enhancement for reference and four variants of Turbo-Twin-HMM speech enhancement. Two different synthesis feature extractions *E1* and *E2* in combination with two different clean speech estimation strategies AP and BP for a total of 588 test files per SNR are shown.

	segmental SNR				PESQ				STOI			
SNR	0 dB	-3 dB	-6 dB	-9 dB	0 dB	-3 dB	-6 dB	-9 dB	0 dB	-3 dB	-6 dB	-9 dB
noisy	1.79	0.40	-1.37	-2.10	1.95	1.69	1.46	1.21	**3.59**	**2.35**	0.35	-0.41
log-MMSE	**2.32**	0.59	-0.56	-1.53	1.90	1.58	1.36	1.06	0.66	0.57	0.49	0.41
E1AP	1.30	0.72	0.09	-0.59	**2.11**	**2.02**	**1.94**	**1.83**	0.68	0.65	0.61	0.57
E1BP	1.30	**0.73**	**0.12**	-0.55	2.02	1.92	1.82	1.71	0.66	0.63	0.59	0.54
E2AP	1.15	0.64	0.04	-0.57	2.08	2.01	1.91	1.82	0.70	0.68	**0.64**	**0.59**
E2BP	1.11	0.63	0.05	**-0.54**	1.99	1.91	1.80	1.68	0.67	0.65	0.60	0.56

We find $i^*(t)$ via a best path search in the state posterior matrix $\tilde{\gamma}$ constrained by the transition structure of the audio model. BP synthesis is done by using only ODFs that belong to states on the best path. Thus, in every frame only a single expectation with weight one is used to calculate an estimate of the clean speech signal, an approach that is more efficient than the MMSE estimator above, albeit at the cost of making a hard decision on the state identity before synthesis.

6. Experimental Setup

We used the GRID corpus [19] as our audio-visual database. In order to obtain noisy mixtures, for each utterance a noise-only segment, with a length equal to that of the clean utterance, was randomly selected from the binaural CHiME noise recordings [20], which contain realistic household noises like steps, washing machines, music, etc., and are hence very challenging due their variety and non-stationarity. To eliminate some low-frequency hum and baseline drift, the noise signal was filtered with an 8-th order Butterworth high-pass filter with cut-off frequency 70Hz, and scaled to yield the desired SNR according to

$$SNR = 10\log_{10}\frac{\sum_{k=0}^{K-1} s_1(k)^2 + s_2(k)^2}{\sum_{k=0}^{K-1} n_1(k)^2 + n_2(k)^2}. \qquad (9)$$

Here, k denotes the sample index. The mean was subtracted from the two channels of the time domain input signals, s_1, s_2, and noise signals n_1 and n_2 before the SNR computation.

The signals were then pre-enhanced using log-MMSE speech estimators according to [21] with beamforming-based noise estimates. Two versions of speech enhancement were compared: On the one hand, we used the log-MMSE estimator given in the source code of Loizou as distributed with [22] in the original parametrization, and on the other hand, we carried out a log-MMSE speech enhancement as described in [23]. The first setup leads to a very high perceptual quality of enhanced speech, whereas the second setup was optimized for best speech recognition performance in the CHiME challenge, where it was first used in [23]. In the following, we refer to the first enhancement by *E1*, and to the second by *E2*.

Finally, REC and SYN features were extracted for the *Twin-HMM*. For REC and SYN features, the framing was chosen equally, so as to ensure temporal consistency. The frame size and frame shift were 20 ms and 10 ms, respectively, and a Hamming window function was applied before carrying out the FFT. The magnitude of this FFT was used as SYN features. In a second step, for the REC features, 13 MFCC coefficients and their first and second time derivative were computed from the STSA features, leading to 39-dimensional REC feature vectors. Video features were extracted exactly as in [14]. Recognition results for these features with our turbo decoding approach and $\lambda_P = 0.1$ are shown in Table 2. λ_T is chosen SNR-dependently between 0.1 and 0.4.

Table 2: Comparison of the recognition accuracy in percent correct for REC-feature variants *E1* (optimized for minimal distortion during synthesis) and *E2* (optimized for best recognition results).

Method	-9 dB	-6 dB	-3 dB	0 dB	∞ dB
E1	87.95%	90.27%	91.13%	93.38%	97.15%
E2	89.67%	91.81%	93.98%	95.34%	98.18%

7. Instrumental Measures

For the evaluation of the speech enhancement system, a range of instrumental quality measures has been applied. We have computed the PESQ measure [24], the segmental SNR, cf. page 45 of [25], was used in the implementation provided in [22], and the short-term objective intelligibility measure (STOI) was used as described in [26]. Mean values of the considered instrumental quality measures for all methods are shown in Table 1.

8. Speech Intelligibility

We have measured the intelligibility of the enhanced speech signals by means of a large-scale listening experiment, using crowd-sourcing tests at CrowdFlower [27]. Each test participant (referred to as a *contributor* in the crowd-sourcing jargon) was asked to transcribe a set of 14 audio signals, covering different signal processing methods, i.e., log-MMSE signal enhancement, *E2BP*, *E2AP*, and *E1AP* as well as an unprocessed noisy signal. For this purpose, we created test sets for 4 different SNR conditions (i.e., -9 dB, -6 dB, -3 dB, and 0 dB), where the SNR was the same for all signals within a given test set. To prevent memorization, we ensured that the same utterance text was only utilized once within a given test set.

Each test set also contained 4 clean utterances that were used for quality control[1]. Only those participants who correctly transcribed at least 75 % of the clean utterances were considered for the experiment.

The transcriptions were recorded using a multiple-choice approach by providing pre-filled selection forms, i.e., radio buttons and drop down menus. Each contributor was allowed to

[1]Quality control can be helpful to identify cheaters (i.e., contributors that are not working fairly) and to exclude contributors that are not sufficiently qualified for a given task (e.g., due to language deficits).

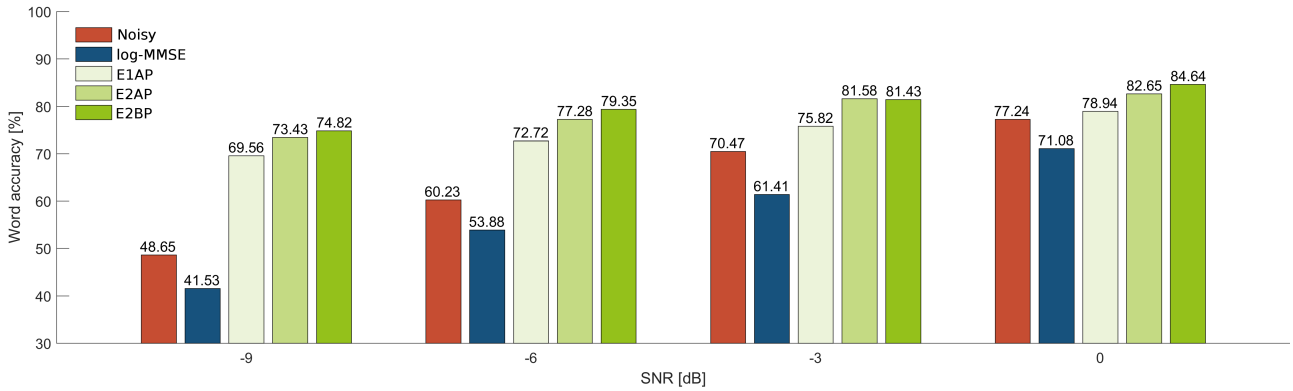

Figure 3: Listening test results. Each score is based on the average results of approximately 260 unique utterances.

participate multiple times but was restricted to solve at most 6 tests with 14 utterances each. We have collected the responses from 690 individual participants, considering only those participants that have passed and maintained the quality control requirements during the test. Overall, we gathered 27.118 transcribed utterances.

Figure 3 shows the results of the listening experiment in terms of the average word accuracy for varying SNRs. There, we can see that the unprocessed noisy signals give a higher word accuracy than the log-MMSE-processed signals, indicating that the latter does not improve the intelligibility of the signals, which is consistent with many previous findings [28]. In contrast, for the proposed approach all variants (i.e., *E2BP*, *E2AP*, and *E1AP*) show improvements in word accuracy for each SNR condition, compared both to the noisy and to the log-MMSE-processed signals. As expected, the results indicate that the Turbo-Twin-HMM is most effective at low SNR conditions. The highest relative performance improvement is observed for *E2BP* at -9 dB, yielding a relative improvement of the word accuracy of 53.79 % and 80.16 % compared to noisy and log-MMSE-enhanced speech, respectively.

The average results across all conditions are shown in Table 3, where an improvement in intelligibility from 64.27% to 80.10% is shown, comparing the noisy signal with the best-performing Turbo-Twin-HMM version, i.e. the *E2BP* setup.

Table 3: Listening test results showing the word accuracy averaged over all SNRs. Each score is based on 1037 utterances.

Noisy	log-MMSE	*E1AP*	*E2AP*	*E2BP*
64.27 %	57.09 %	74.30 %	78.78 %	80.10 %

9. Discussion

When comparing the results of the instrumental quality measures and those of the listening tests, it is of note that high values in the instrumental measures do not correspond to high values of intelligibility. Whereas at high SNRs, noisy or log-MMSE-enhanced speech leads to good values especially of the STOI measure, the Turbo-Twin-HMM results are the most intelligible signals in all cases. This seeming discrepancy is, however, not so surprising if one considers the fact that our approach actually synthesizes the speech spectra from a clean-speech model. While this does introduce greater changes in the signal form, and hence deviations from the clean-speech reference, and thus may negatively impact intelligibility estimates such as the STOI measure, indeed the signal does gain in intelligibility, as missing acoustic information is replaced by means of the available audio-visual speech model. This is most apparent in the results of the *E2BP* approach, which improves intelligibility from 46.6% to 74.8% at -9dB, and still yields considerable improvements of intelligibility at 0dB SNR, going from 77.2% to 84.6%.

Another point of interest is the comparison between the two log-MMSE pre-enhancement settings. Whereas the setting *E1*, optimized for human listening, typically leads to better PESQ and segmental-SNR values, due to our ASR-model-based approach, the best listening test results are still attained using *E2*, the approach optimized for automatic speech recognition performance.

Among the two tested fusion approaches, best-path synthesis is the clear favorite in terms of intelligibility, even though the MMSE estimate employed in AP is typically superior regarding instrumental measures. Again, this points to an interesting fact—as noted in [29], the use of reference-based methods for estimating speech intelligibility is questionable in many model-based speech enhancement schemes, and should be replaced by listening tests, or, ideally, by yet-to-be-developed intelligibility assessment approaches based on more complex models of speech.

10. Conclusions

We have introduced the Turbo-Twin-HMM, a new model for audio-visual speech enhancement. Combining the capability of the *Twin-HMM* for multi-modal speech enhancement with the recognition accuracy of turbo decoding, the approach has been successful in clearly improving speech intelligibility in highly noisy environments. Its use is most effective at strong negative SNRs, where, based on the information of the video channel, missing data in the acoustic domain can actually be recovered by utilizing a state-dependent model of the clean speech spectral amplitudes. The efficacy of the approach has been demonstrated not only in terms of instrumental measures of speech quality and intelligibility, but also by large-scale listening tests that confirm the expected notable improvements of intelligibility.

Whereas here, we have concentrated on a proof-of-concept system based on the GRID corpus, in future work, it will be interesting to apply this framework for large vocabularies and to extend its applications to de-reverberation and source separation, with the goal of employing it, e.g. in mobile devices, for video-assisted speech enhancement.

11. References

[1] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Using Twin-HMM-based audio-visual speech enhancement as a front-end for robust audio-visual speech recognition," in *Proc. Interspeech*, Lyon, France, August 2013.

[2] S. Shivappa, B. D. Rao, and M. M. Trivedi, "Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition," in *Proc. ICASSP*, 2008, pp. 2241–2244.

[3] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Johns Hopkins University, CLSP, Tech. Rep. WS00AVSR, 2000.

[4] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Twin-HMM-based audio-visual speech enhancement," *Proceedings ICASSP*, pp. 3726–3730, 2013.

[5] A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. ICASSP*, vol. 2, 2002, pp. 2013–2016.

[6] R. Astudillo, D. Kolossa, and R. Orglmeister, "Propagation of statistical information through non-linear feature extractions for robust speech recognition," in *Proc. MaxEnt2007*, 2007.

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

[8] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 19, no. 4, pp. 822–836, 2011.

[9] R. M. Nickel, R. F. Astudillo, D. Kolossa, S. Zeiler, and R. Martin, "Inventory-style speech enhancement with uncertainty-of-observation techniques," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 3877–3880.

[10] R. M. Nickel, R. F. Astudillo, D. Kolossa, and R. Martin, "Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing," *IEEE Trans. Audio, Speech & Language Processing*, vol. 21, no. 5, pp. 983–997, 2013.

[11] D. Kolossa, R. Nickel, S. Zeiler, and R. Martin, "Inventory-based audio-visual speech enhancement," in *Proc. Interspeech*, Portland, OR, September 2012.

[12] C. W. Seymour and M. Niranjan, "An HMM based cepstral-domain speech enhancement scheme," in *Proc. Interspeech*, Yokohama, Japan, 1994, pp. 1595–1598.

[13] S. Receveur, R. Weiss, and T. Fingscheidt, "Turbo automatic speech recognition," *Audio, Speech, and Language Processing (TASLP), IEEE/ACM Transactions on*, vol. 99, pp. 1–1, 2016.

[14] S. Zeiler, R. Nickel, N. Ma, G. J. Brown, and D. Kolossa, "Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis," in *ICASSP*. IEEE, 2016, pp. 1–2.

[15] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in *Proc. ICC*, vol. 2, Geneva, 1993, pp. 1064–1070.

[16] C. Berrou and A. Glavieux, "Near optimum error-correcting coding and decoding: Turbo Codes," *IEEE Transactions on Communications*, vol. 44, no. 10, pp. 1261–1271, Oct. 1996.

[17] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 23, no. 5, pp. 863–876, 2015.

[18] S. Gergen, S. Zeiler, A. Hussen Abdelaziz, R. Nickel, and D. Kolossa, "Dynamic stream weighting for turbo-decoding-based audiovisual ASR," *accepted for publication at Interspeech 2016*.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[20] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 443–445, Apr. 1985.

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

[23] D. Kolossa, R. F. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. Neto, and R. Martin, "CHIME challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques," in *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, 2011, pp. 6–11.

[24] 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," *Proc. ICASSP*, vol. 2, pp. 749–752, 2001.

[25] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Prentice-Hall, 1988.

[26] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214 – 4217.

[27] CrowdFlower, Inc, "CrowdFlower," as of February 2016, http://www.crowdflower.com.

[28] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, Jan 2011.

[29] T. Fingscheidt and P. Bauer, "A phonetic reference paradigm for instrumental speech quality assessment of artificial speech bandwidth extension," in *Proc. of PQS Workshop 2013, Vienna, Austria*, 2012.