Learning a concatenative resynthesis system for noise suppression

Michael I. Mandel
Computer Science & Engineering
The Ohio State University
Columbus, OH, 43210
mandelm@cse.ohio-state.edu

Young Suk Cho
Computer Science & Engineering
The Ohio State University
Columbus, OH, 43210
cho.648@osu.edu

Yuxuan Wang
Computer Science & Engineering
The Ohio State University
Columbus, OH, 43210
wangyuxu@cse.ohio-state.edu

***Abstract*—This paper introduces a new approach to dictionary-based source separation employing a learned non-linear metric. In contrast to existing parametric source separation systems, this model is able to utilize a rich dictionary of speech signals. In contrast to previous dictionary-based source separation systems, the system can utilize perceptually relevant non-linear features of the noisy and clean audio. This approach utilizes a deep neural network (DNN) to predict whether a noisy chunk of audio contains a given clean chunk. Speaker-dependent experiments on the small-vocabulary CHiME2-GRID corpus show that this model is able to accurately resynthesize clean speech from noisy observations. Preliminary listening tests show that the system's output has much higher audio quality than existing parametric systems trained on the same data, achieving noise suppression levels close to those of the original clean speech.**

***Index Terms*—Speech; noise suppression; concatenative synthesis; corpus-based; nonparametric**

I. Introduction

Acoustic noise is very disruptive to voice communication. Even mild levels of noise reduce the intelligibility of speech for the hearing impaired and for hearing aid users [1]. They also increase cognitive load [2] in both normal hearing and hearing impaired listeners. Current noise suppression technologies typically attempt to modify the noisy signal to make it sound more like the original clean signal. We propose here instead replacing the noisy signal with a signal created by concatenative synthesis [3]. Such an approach can almost entirely eliminate the noise while maintaining the high speech quality that concatenative synthesizers have shown. Such a system could be fielded in a mobile phone that records its owner's voice in a quiet environment and then resynthesizes new speech from those recordings, even in noisy environments.

To perform this concatenative synthesis, we propose learning a non-linear similarity function between pairs of "chunks" of noisy and clean speech such that pairs involving the same speech are given high similarity while pairs involving different speech are not. At test time, this function will be used in conjunction with a corpus of clean speech "units" to resynthesize the clean speech. In particular, clean dictionary elements will be selected to replace noisy observations such that they are similar under our learned function and such that adjacent dictionary elements are compatible with one another.

There are existing noise suppression algorithms based on non-negative matrix factorization [4] and sparse coding [5] that model noisy speech as a linear sum of speech and noise bases. These models are typically applied in the Fourier magnitude and power spectrum domains, where additivity only approximately holds. Such models cannot be applied to perceptually motivated features like the log mel spectrum [6], because additivity does not hold there. Such domains require a learned, non-linear metric, such as the one presented here.

II. Related work

Several methods have been proposed recently for corpus-based speech enhancement [7–11]. Work in one research thread [7–9] builds dictionaries of noisy speech, adding noise to known clean speech. While this makes comparing the observation to the dictionary simple, it requires a very large dictionary and access to the noise at training time. A second research thread [10, 11] builds dictionaries out of noise-suppressed speech. It performs standard noise suppression on synthetic mixtures at training time and on the observation at test time, and then performs matching in this noise suppressed domain. While more robust and generalizable than the noise-dependent approach, this method is still sensitive to the performance of the noise suppressor. Such an approach could be applied in tandem with our proposed approach.

There are several non-linear metric learning algorithms [12–14] and several learning to rank algorithms [15–18] that have been applied to other problems but are relevant to this one as well. The closest are perhaps [19, 20], which learn neural networks to map examples into a space in which pairs of examples from the same class are close to each other and pairs from different classes are far apart. Our system differs from those in that we process the dictionary elements and noisy observations through a single combined network, allowing it to learn different transformations for each signal type and interactions between them. By predicting a similarity score instead of a ranking, we can also combine our predictions with transition scores to improve temporal continuity.

III. Non-linear dictionary-based denoising

Figure 1 shows an illustration of the use of a learned non-linear metric for audio denoising. It shows that an incoming noisy audio stream is decomposed into temporally-overlapping

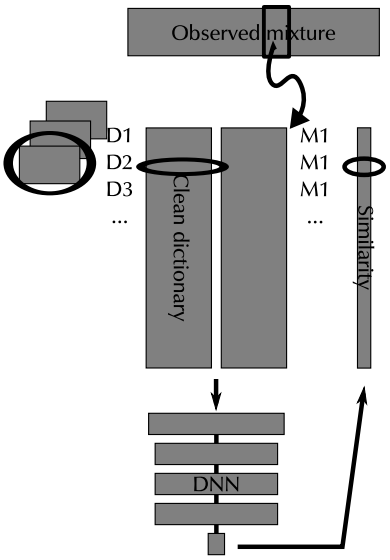

Fig. 1. Overview of denoising system. Each "chunk" of the observed mixture signal is paired with each "chunk" of clean speech in the dictionary and input to a deep neural network (DNN), providing a similarity score. The noisy chunk can be replaced by the dictionary element with the highest similarity score.

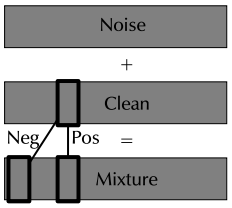

Fig. 2. Two pairs of clean and noisy chunks illustrating the collection of training data for the paired-input DNN. In the positive pair, the clean speech chunk is present in the noisy mixture chunk; in the negative pair it is not.

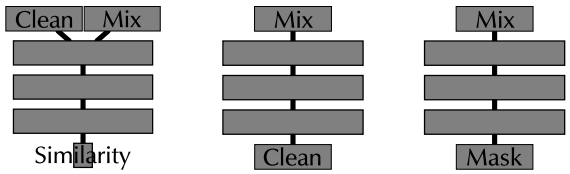

Fig. 3. Structure of the DNNs used in the experiments. Predicting (a) similarity from a (clean, noisy) pair of chunks, (b) clean chunk from noisy chunk, (c) ideal ratio mask from noisy chunk.

chunks. Each chunk is then compared with every element in the dictionary using a learned similarity function, to create a matrix of similarity values. The best path through this matrix is then found, incorporating a transition affinity between dictionary elements and the audio corresponding to the dictionary elements on the best path is assembled into the output. Because the dictionary only consists of clean speech, the resynthesis contains no noise aside from incorrect or misplaced dictionary elements. Mathematically, the best path is

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \prod_t p(z_t = j \mid x_t) p(z_t = j \mid z_{t-1} = i) \quad (1)$$

$$= \underset{\mathbf{z}}{\operatorname{argmax}} \prod_i g(z_j, x_i) T_{ij} \quad (2)$$

where, $g(z_j, x_i)$ is the similarity function between clean chunk z_j and noisy chunk x_i defined in Section III-A and T_{ij} is the transition matrix defined in Section III-B.

Once the best path through the dictionary is found, the clean speech is resynthesized using an overlap-add technique. Specifically, the audio corresponding to each selected dictionary chunk is extracted and added to an output buffer with a linear fade-in and fade-out of 16 ms. The overlap between query chunks can be anything up to the length of a chunk, but we have found that an overlap of half of a chunk provides a good compromise between continuity and computational cost.

A. Learning the similarity function

We thus endeavor to learn a non-linear similarity function, $g(z, x)$ between clean speech chunks, $\{z_j\}_{j=1}^J$, and chunks of noisy speech, $\{x_i\}_{i=1}^I$. In the time domain, these chunks comprise a few hundred milliseconds of audio. Input audio for both models is first transformed into the perceptually motivated log mel spectrum [6]. We assume that each noisy chunk is made up of a sum of at least one dictionary chunk and extraneous noise from outside the dictionary. Thus we define y_{ij}, the desired output of our similarity function, to be 1 when x_i contains z_j and 0 otherwise. See Figure 2 for an illustration of how this ground truth target data is derived from speech mixed with noise.

In this work, we use a deep neural network (DNN) as $g(z, x)$. The structure of the DNN is shown in Figure 3(a). It takes as input 11 frames of the log mel spectrograms of a clean and noisy chunk, rasterized and concatenated together, making a total of 484 dimensions. It has 4 hidden layers of 1024 rectified linear units (ReLU) [21] each, and a single sigmoid output unit predicting y_{ij}. It is initialized with random weights and trained using backpropagation with dropout of 20% of the hidden units [22] and ADA-Grad stochastic gradient descent [23]. Training minimizes the cross-entropy between y_{ij} and $g(z_j, x_i)$

$$\sum_{i,j} -y_{ij} \log g(z_j, x_i) - (1 - y_{ij}) \log(1 - g(z_j, x_i)). \quad (3)$$

This ground truth is perhaps more restrictive than it needs to be. Future work will explore the inclusion of inexact matches based on various characterizations of the dictionary elements.

B. Transition affinities

We also compute, $T(z_i, z_j)$, the affinity of transitioning from dictionary element z_i to dictionary element z_j. For now, this computation is deterministic and not learned. We construct a matrix T_{ij} of all possible such transitions by comparing the log mel spectrum of the end of z_i to that of the beginning of z_j. In particular

$$T_{ij} = \exp\left(-d_\tau(z_i, z_j)/\gamma\right) \quad (4)$$

where $d_\tau(z_i, z_j)$ is the Euclidean distance between the log mel spectra of the last τ frames of z_i and the first τ frames of z_j and γ is a parameter that controls the mapping of distances to affinities. Because our dictionaries contain approximately 60,000 elements, computing the whole transition matrix is feasible, but storing it is not. We thus store only the largest 10,000,000 (0.28% of) entries and set all others to be equal to the smallest stored entry. Although we do not take advantage of the fact here, computation of these affinities could be performed on different features than are used by the DNN.

IV. Experiments

We use the data from the CHiME2-GRID small vocabulary dataset [24]. The Second Computational Hearing in Multisource Environments challenge (CHiME2) simulates read speech in a living room environment with recorded household noises at six different signal to noise ratios (SNRs) from -6 dB to 9 dB. It utilizes sentences from the GRID corpus [25], which consist of six short words of the form, "⟨*command*⟩ ⟨*color*⟩ ⟨*preposition*⟩ ⟨*letter*⟩ ⟨*digit*⟩ ⟨*adverb*⟩", for example, "place blue at A 9 again." In assembling the data, the authors selected noises that could achieve each SNR with only minor adjustments in gain, creating more realistic mixtures. The noises are mainly the speech of women and children, music, and various household activities. We use clean spatialized speech from the "reverberant" condition and we use speech and noise mixtures from the "isolated" condition, which are cropped to the duration of the speech. Although all signals are provided in stereo, we average the two channels together to perform monaural computations.

We use the official split of the mixtures into 500 training utterances from each speaker at various SNRs and the same 24 development utterances (used for testing) at each of the six SNRs. Each sentence is approximately two seconds long, for a total of 16 minutes of training speech for each speaker. The training and development sets contain both different utterances and different noises.

From these utterances, which are sampled at 16 kHz, we compute the log mel spectrogram in Matlab using [26], with an FFT frame size of 32 ms and a hop size of 16 ms. We found that chunks of 11 frames, corresponding to 192 ms total, strike a good balance between generality and meaningfulness. For training, we extract 11-frame chunks from the log mel spectrogram that overlap by 10 frames with their neighbors. This leads to a training dataset of 124,080 clean-noisy pairs of chunks, half matching and half not matching. The testing dictionary is constructed from all 500 training utterances, and contains 67,040 clean chunks. Our experiments are speaker-dependent, in that models are trained and tested on different utterances from the same speaker. We selected speaker 3, a male, because his speech had the longest duration.

A. Comparison systems

We compare our model with two mask-based source separation systems using similar neural networks trained on the matching half of the training dataset. The first, shown in Figure 3(b), predicts the log mel spectrum of the clean speech from that of the noisy speech, and the second, shown in Figure 3(c) predicts the ideal ratio mask [27] from the noisy speech. Both thus take 242 input dimensions and output 242 predictions. Both networks use four hidden layers of 1024 ReLU units each. The first uses linear output units to predict the log mel spectrum of the clean speech with the mean squared error criterion. Its estimates of the clean speech are then used to construct a Wiener-like mask by subtracting the log mel spectrum of the noisy mixture from it and converting to power spectrum. The second network uses sigmoid output units with the cross-entropy error criterion with the ideal ratio mask. Its predictions (continuous values between 0 and 1) are then used directly to mask the noisy mixture.

TABLE I
Retrieval metrics for correct dictionary element averaged over 500 noisy query chunks selected at random from the CHiME2-GRID devel set. For average rank, lower is better.

	Euclidean	Paired NN
Number of dictionary chunks	2899	2899
Average rank of correct chunk	630	5
Precision-at-1	21.5%	72.3%

B. Ranking dictionary elements

We first evaluate the performance of our system by measuring the retrieval performance of the dictionary element ranking system. We do so here using a dictionary in which exactly one element is the ground truth speech that was used to generate the query chunk and the other dictionary elements were not. We then run the network on inputs consisting of the pairing of each dictionary chunk with the noisy query chunk and measure the rank of this single correct answer in the list of all network outputs. We also measure the proportion of the time that the correct dictionary element is the top ranked in this list. We compare the performance of the paired-input neural network with a simple Euclidean distance. This experiment is performed with 500 randomly selected noisy chunks as queries and these metrics are averaged over all trials. The results of this experiment can be seen in Table I. They show that the paired neural network performs significantly better than the Euclidean distance in retrieving the correct dictionary element for each noisy query chunk. Not only does the correct chunk rank very close to the top of the list on average, it is the top element for 78.3% of the queries.

C. Denoising signals

In an actual application of the proposed system, there is very little chance of the exact speech that went into a mixture being present in the dictionary. Furthermore, there may be several dictionary elements that are very close to one another, any of which would make a good replacement for the noisy speech. Thus, to evaluate the system more realistically, we resynthesize the clean speech from a noisy utterance and measure the subjective quality and intelligibility of the result. For each clean utterance in the CHiME2-GRID development

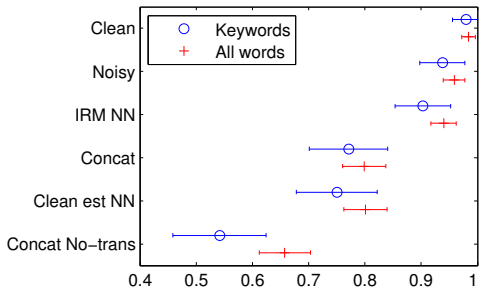

Fig. 4. Intelligibility subjective listening test results showing the average percentage correctly identified of all words and just the keywords (letters and numbers) for each system. Errors bars show 95% confidence intervals.

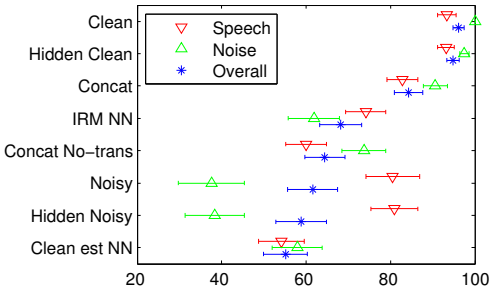

Fig. 5. MUSHRA listening test results showing the average subjective speech quality, noise suppression, and overall quality for each system. Higher numbers are better in all cases. Errors bars show 95% confidence intervals.

set, we randomly selected a mixture at one SNR to process and evaluate[1]. Processing a 2 second mixture using the proposed system took approximately 8 CPU-minutes on an Intel Core i5-2500 system.

The intelligibility test compared six different versions of each mixture: the original clean speech (Clean), the unprocessed mixture (Noisy), the concatenative resynthesis without transition information (Concat No-trans), the concatenative resynthesis with transition information (Concat), and the neural networks described in Section IV-A trained to predict the ideal ratio mask (IRM NN) and the clean speech (Clean est NN) from the noisy speech. Three subjects evaluated these files, two of whom were native English speakers. They listened to all 24 mixtures processed by all 6 systems in different random orders. Subjects were given a copy of the GRID grammar and asked to transcribe the sentences as best they could, noting that they did not necessarily have to adhere to the grammar. The GRID task itself only measures the accuracy on the letter and number in each sentence, so we report this number as well.

The results of the intelligibility test for each system averaged over all files are shown in Figure 4, which plots both the accuracy on all words and the accuracy on the letter and number keywords. The same trends are evident in both accuracies: intelligibility is already high for the noisy speech, there is no processing that increases the intelligibility significantly, the concatenative synthesis has the worst intelligibility, but it is significantly improved by the use of the transition matrix.

The speech quality test compared the same six systems under a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) paradigm. For each mixture, listeners were presented with the reference clean and noisy speech, and then with the same six systems as above unlabeled and in a random order. The comparisons were presented in a different random order for each listener. They were asked to rate each processed mixture in terms of speech quality, noise suppression, and overall quality each on a scale from 0 (poor) to 100 (excellent). They were also instructed that an utterance could have high quality while having low intelligibility if it sounded like natural speech. Four listeners participated, two of whom had never heard any of the processed mixtures before and two who had heard some of them. Two finished all 24 comparisons, one finished 19, and one finished 10.

The results of the speech quality test for each system are shown in Figure 5, averaged across files and listeners. These results show that the concatenative resynthesis system produces speech with the highest quality in all three categories, with the noise suppression quality approaching that of the original clean speech. These differences are significant for all three categories. We noticed a large discrepancy between these subjective results and both the PEASS prediction of audio quality [28] and the STOI prediction of intelligibility [29]. We believe that this is because these metrics measure similarity to a single reference and while our system produces natural and high quality speech, it differs from this reference.

V. Conclusions

This paper has introduced a new paradigm for corpus-based speech enhancement, namely the use of a learned non-linear similarity function to make noise-robust predictions of the dictionary elements that could explain a noisy speech utterance. Because it is non-linear and learned, this function is able to use arbitrary, perceptually motivated features as input. When combined with a transition similarity function it becomes essentially a concatenative resynthesis system. It outputs speech that is significantly higher quality than competing parametric systems trained on the same data, although of lower intelligibility. In the future, numerous extensions to the system are possible for performing nonparametric speech recognition, pitch tracking, etc., by annotating the clean dictionary elements before matching them to observations. The computational demands of using significantly larger dictionaries, as would be required for large vocabulary tasks, could be reduced using a beam search in the Viterbi decoding or by using some form of locality sensitive hashing to identify candidate dictionary matches for each noisy chunk in constant time.

Acknowledgments: This work was supported in part by a Google Faculty Research Award and performed while the first author was visiting Télécom ParisTech. The authors would like to thank Brian Kulis, Gaël Richard, and Slim Essid for useful conversations and Johanna Devaney, Clément Laroche, and Paul Magron for their assistance.

[1]These files can be heard at http://mr-pc.org/work/globalsip14/

REFERENCES

[1] J. I. Alcántara, B. C. J. Moore, V. Kühnel, and S. Launer. "Evaluation of the noise reduction system in a commercial digital hearing aid: Evaluación del sistema de reducción de ruido en un auxiliar auditivo digital comercial". In: *Intl. J. Audiol* 42.1 (2003), pp. 34–42.

[2] A. A. Zekveld, S. E. Kramer, and J. M. Festen. "Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response". In: *Ear Hear.* 32.4 (2011), pp. 498–510.

[3] A. J. Hunt and A. W. Black. "Unit selection in a concatenative speech synthesis system using a large speech database". In: *Proc. IEEE Intl. Conf. Acous. Speech & Sig. Proc.* Vol. 1. IEEE, 1996, pp. 373–376.

[4] D. D. Lee and H. S. Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (1999), pp. 788–791.

[5] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies. "Sparse Representations in Audio and Music: From Coding to Source Separation". In: *Proc. IEEE* 98.6 (2010), pp. 995–1005.

[6] S. Imai. "Cepstral analysis synthesis on the mel frequency scale". In: *Proc. IEEE Intl. Conf. Acous. Speech & Sig. Proc.* Vol. 8. IEEE, 1983, pp. 93–96.

[7] X. Xiao and R. M. Nickel. "Speech Enhancement With Inventory Style Speech Resynthesis". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 18.6 (2010), pp. 1243–1257.

[8] J. Ming, R. Srinivasan, and D. Crookes. "A Corpus-Based Approach to Speech Enhancement From Nonstationary Noise". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 19.4 (2011), pp. 822–836.

[9] A. Ogawa, K. Kinoshita, T. Hori, T. Nakatani, and A. Nakamura. "Fast segment search for corpus-based speech enhancement based on speech recognition technology". In: *Proc. IEEE Intl. Conf. Acous. Speech & Sig. Proc.* 2014.

[10] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, et al. "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation". In: *Proc. 1st Intl. Workshop on Machine Listening in Multisource Environments (CHiME)*. 2011, pp. 12–17.

[11] R. M. Nickel, R. F. Astudillo, D. Kolossa, and R. Martin. "Corpus-Based Speech Enhancement With Uncertainty Modeling and Cepstral Smoothing". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 21.5 (2013), pp. 983–997.

[12] P. Jain, B. Kulis, J. Davis, and I. Dhillon. "Metric and kernel learning using a linear transformation". In: *J. Mach. Learn. Res.* 13 (2012), pp. 519–547.

[13] L. Torresani and K. Lee. "Large margin component analysis". In: *NIPS*. 2007, pp. 1385–1392.

[14] K. Weinberger and L. Saul. "Fast solvers and efficient implementations for distance metric learning". In: *Proc. ICML*. 2008, pp. 1160–1167.

[15] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. "Learning to Rank Using Gradient Descent". In: *Proc. ICML*. 2005, pp. 89–96.

[16] C. J. Burges. *From RankNet to LambdaRank to LambdaMART: An Overview*. Tech. rep. MSR-TR-2010-82. Microsoft Research, 2010.

[17] B. McFee and G. R. Lanckriet. "Metric learning to rank". In: *Proc. ICML*. 2010, pp. 775–782.

[18] D. Lim and G. Lanckriet. "Efficient Learning of Mahalanobis Metrics for Ranking". In: *Proc. ICML*. 2014.

[19] S. Chopra, R. Hadsell, and Y. LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *Proc. IEEE CVPR*. Vol. 1. Los Alamitos, CA, USA: IEEE, 2005, 539–546 vol. 1.

[20] R. Salakhutdinov and G. E. Hinton. "Learning a nonlinear embedding by preserving class neighbourhood structure". In: *Proc. AISTATS*. 2007, pp. 412–419.

[21] V. Nair and G. E. Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proc. ICML*. 2010, pp. 807–814.

[22] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (2012).

[23] J. Duchi, E. Hazan, and Y. Singer. "Adaptive subgradient methods for online learning and stochastic optimization". In: *J. Mach. Learn. Res.* 12 (2011), pp. 2121–2159.

[24] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni. "The second 'CHiME' Speech Separation and Recognition Challenge: Datasets, tasks and baselines". In: *Proc. IEEE Intl. Conf. Acous. Speech & Sig. Proc.* 2013, pp. 126–130.

[25] M. Cooke, J. Barker, S. Cunningham, and X. Shao. "An audio-visual corpus for speech perception and automatic speech recognition". In: *The Journal of the Acoustical Society of America* 120.5 (2006), p. 2421.

[26] D. P. W. Ellis. *PLP and RASTA (and MFCC, and inversion) in Matlab*. 2005.

[27] S. Srinivasan, N. Roman, and D. Wang. "Binary and ratio time-frequency masks for robust speech recognition". In: *Speech Comm.* 48.11 (2006), pp. 1486–1501.

[28] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. "Subjective and Objective Quality Assessment of Audio Source Separation". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 19.7 (2011), pp. 2046–2057.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 19.7 (2011), pp. 2125–2136.