Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm

Lars Bramsløw, Gaurav Naithani, Atefeh Hafez, Tom Barker, Niels Henrik Pontoppidan, and Tuomas Virtanen

Citation: The Journal of the Acoustical Society of America **144**, 172 (2018); doi: 10.1121/1.5045322
View online: https://doi.org/10.1121/1.5045322
View Table of Contents: http://asa.scitation.org/toc/jas/144/1
Published by the Acoustical Society of America

Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm[a)]

Lars Bramsløw,[1,b)] Gaurav Naithani,[2] Atefeh Hafez,[1] Tom Barker,[2,c)]
Niels Henrik Pontoppidan,[1] and Tuomas Virtanen[2]
[1]*Eriksholm Research Centre, Oticon A/S, Rørtangvej 20, DK-3070 Snekkersten, Denmark*
[2]*Tampere University of Technology, Laboratory of Signal Processing, Tampere, P.O. Box 553, FI-33101 Tampere, Finland*

(Received 23 November 2017; revised 18 May 2018; accepted 18 June 2018; published online 11 July 2018)

Hearing aid users are challenged in listening situations with noise and especially speech-on-speech situations with two or more competing voices. Specifically, the task of attending to and segregating two competing voices is particularly hard, unlike for normal-hearing listeners, as shown in a small sub-experiment. In the main experiment, the competing voices benefit of a deep neural network (DNN) based stream segregation enhancement algorithm was tested on hearing-impaired listeners. A mixture of two voices was separated using a DNN and presented to the two ears as individual streams and tested for word score. Compared to the unseparated mixture, there was a 13%-point benefit from the separation, while attending to both voices. If only one output was selected as in a traditional target-masker scenario, a larger benefit of 37%-points was found. The results agreed well with objective metrics and show that for hearing-impaired listeners, DNNs have a large potential for improving stream segregation and speech intelligibility in difficult scenarios with two equally important targets without any prior selection of a primary target stream. An even higher benefit can be obtained if the user can select the preferred target via remote control.
 https://doi.org/10.1121/1.5045322

[JFL] Pages: 172–185

I. INTRODUCTION

Competing voices is a commonly encountered listening challenge for a hearing aid user, e.g., during family dinners, while watching TV, social events, etc. It is well known that hearing-impaired individuals struggle when being in noisy situations, and the well-known "cocktail party problem" (Cherry, 1953) is a major challenge for them. In fact, with hearing loss, a situation as simple as two competing voices next to each other across a table causes much informational masking (Ezzatian *et al.,* 2015; Ihlefeld and Shinn-Cunningham, 2008), causing both voices to mask and disturb one another (Brungart, 2001). A hearing-impaired person may benefit from moderate spatial separation between the two talkers, but will still be affected by the informational masking (Neher *et al.,* 2007). In many cases, the two mutually interfering talkers are too close for the hearing-impaired person to segregate them sufficiently. The challenge for a hearing aid algorithm is then effectively reduced to a single channel problem.

Single channel two-talker speech separation has been an active research area for several decades and various types of methods have been employed to address it. It can broadly be divided into three approaches: signal processing-based methods, model-based methods, and supervised learning-based methods. An early example of signal processing-based method is the harmonic selection principle described by Parsons (1976), by which two talkers are separated using their different fundamental frequency (pitch) to separate out the two harmonic structures. No formal listening test was conducted. The algorithm was later enhanced and evaluated by Stubbs and Summerfield (1990) using also a cepstral technique to do a pitch-based separation. In a listening test with natural spoken sentences and normal pitch variation ("intoned sentences"), they found approximately a 30%-point benefit for the target sentences in normal-hearing listeners and no benefit for hearing-impaired listeners. Spectral subtraction (Boll, 1979) based techniques for two-talker separation have been employed by Hanson and Wong (1984) and Naylor and Boll (1987). A method based on sinusoidal modeling of speech was utilized by Quatieri and Danisewicz (1990).

Other approaches include model-based approaches, e.g., Roweis (2001), who demonstrated separation of male and female voices based on prior learning of the clean voices using hidden Markov models. Pontoppidan and Dyrholm (2003) investigated the computational complexity, Bach and Jordan (2005) investigated the requirements of knowing the voices beforehand, and in a series of studies Wang investigated segregation of voiced and unvoiced speech parts based on pitch (Roman and Wang, 2006; Wang and Hu, 2006). Other notable model-based approaches include basis

[a)]Different portions of this work were presented at the Challenges in Hearing Aid Technology (CHAT) Workshop, Stockholm, Sweden, August 19, 2017 and International Hearing Aid Conference (IHCON), Lake Tahoe, CA, USA, August 10–14, 2016.
[b)]Electronic mail: labw@eriksholm.com
[c)]Present address: Cirrus Logic, London, UK.

decomposition methods, e.g., application of independent component analysis (ICA) to monaural separation (e.g., Jang and Lee, 2004), non-negative matrix factorization-based methods (e.g., Virtanen, 2007), and latent variable decomposition methods (e.g., Raj and Smaragdis, 2005).

Supervised learning-based methods in the context of speech denoising have been around for decades, framed as a regression problem utilizing shallow neural networks to predict clean speech spectrum from mixture spectrum (Tamura and Waibel, 1988; Xie and Van Compernolle, 1994). More recently, drawing inspiration from computational auditory scene analysis (CASA) (Wang and Brown, 2006), supervised speech separation was framed as having the goal of estimating a time-frequency mask (binary or soft) (e.g., Han and Wang, 2012; Seltzer *et al.,* 2000). It has been reported that significant improvement in speech intelligibility can be achieved both for normal hearing and hearing-impaired listeners with ideal binary masking (Wang, 2008; Wang *et al.,* 2009). With the advent of deep neural networks (DNNs), a large improvement in the performance of supervised speech separation has been reported starting with Wang and Wang (2013). Various network architectures have been employed, e.g., feedforward DNNs (Grais *et al.,* 2014; Xu *et al.,* 2015), recurrent neural networks (Erdogan *et al.,* 2015; Huang *et al.,* 2015; Weninger *et al.,* 2014), deep autoencoders (Lu *et al.,* 2013), convolutional neural networks (Chandna *et al.,* 2017; Park and Lee, 2016), convolutional recurrent neural networks (Naithani *et al.,* 2017), etc. These DNN-based approaches have employed either time-frequency masking (Huang *et al.,* 2015; Weninger *et al.,* 2014; Williamson and Wang, 2017) or spectral mapping (Grais *et al.,* 2014; Park and Lee, 2016; Xu *et al.,* 2014, 2015) approaches. A more comprehensive discussion of DNN based supervised speech separation can be found from Wang and Chen (2017). In low latency scenarios, DNN based speech separation has been reported by Naithani *et al.* (2016) and Naithani *et al.* (2017) for algorithmic delay <10 ms. Recently, a time domain DNN approach has been proposed by Luo and Mesgarani (2017), where an algorithmic latency of approximately 5 ms was reported.

In recent years, advanced source separation algorithms using deep neural networks have been successfully applied to separation of competing voices and to the speech in noise problem as a noise reduction algorithm. The first benefits for people with hearing impairment came in 2013, where Healy *et al.* (2013) investigated the benefit in speech-shaped noise and with babble at various signal-to-noise ratios (SNRs). Testing both normal-hearing and hearing-impaired listeners indicated that processing by the algorithm increased the intelligibility in all conditions. These improvements were larger for hearing-impaired (HI) listeners, given their poorer baseline performance, and especially for the modulated background, and for the lowest SNRs. Substantial benefits were reported, allowing several HI listeners to improve word recognition scores from near zero to values above 70%. Note that most of the separation algorithms published so far have processing delays that would be above the typical delays in hearing aids of 5–10 ms, as found to be preferred by hearing impaired listeners (Bramsløw, 2010).

The benefits of single-talker separation from babble and stationary noise using DNN and different types of time-frequency masks were explored by Wang (2015), who found DNNs to outperform classical source separation techniques and to provide benefits of 50%-points word score improvement for hearing-impaired (HI) listeners, the same results were published earlier by Healy *et al.* (2013). The problem of a single competing talker was investigated by Healy *et al.* (2017), who found an average benefit of 59%-points for HI listeners at the least favorable SNR of -9 dB. The mean intelligibility achieved by the HI listeners using the algorithm was equivalent to that of young NH listeners without processing, under conditions of identical interference.

An important feature of a separation algorithm is what degree of generalization it possesses: achieving high scores on known speech and noise material is only interesting if this generalizes to, e.g., new noise segments, new speech from the same talker and possibly new talkers, e.g., Kolbæk *et al.* (2017). This is not a concern in the present study: here it is assumed that the two voices to be separated are known by the algorithm, so the deep neural network must have access to isolated segments of speech from both voices, which in real life can be obtained by recording and storing segments with each voice in isolation.

It is important to find the most relevant way of assessing the benefit from such separation algorithms, and even though many studies use objective metrics, such as the STOI and ESTOI (Jensen and Taal, 2016; Taal *et al.,* 2011), the final assessment should be a speech recognition test on the target group of listeners. Earlier work on separation has used consonant recognition in hearing-impaired listeners as an outcome measure and found a benefit in both speech shaped and babble noises (Healy *et al.,* 2014). This type of benefit was also confirmed in novel noise types, which is a basic requirement for successful application of such an algorithm (Healy *et al.,* 2015).

When assessing the benefit of noise reduction and other signal enhancement algorithms, the traditional choice for hearing-impaired listeners would be sentence tests with natural sentences, such as the Hearing in Noise Test (HINT) (Nilsson *et al.,* 1994) or matrix-type structure sentences such as the German OLSA (Kollmeier and Wesselkamp, 1997) and the Danish Dantale2 (Wagener *et al.,* 2003). In these tests, there is a designated target sentence and a designated noise type, typically playing continuously in the background. The masker can be stationary noise, babble or even a single talker.

Concurrent talkers/voice-on-voice/competing voices scenarios have also been reported in the literature. One such example is the CRM test in English using simultaneous talkers with the same sentence structure: names for cueing and colors plus numbers for response options (Bolia *et al.,* 2000). The fixed and time-aligned sentence structure of the test is suited for exploring low-level spatial and phonemic cues, but the sentences do not represent ordinary conversations. Helfer *et al.* (2010) also investigated competing voice disturbances, but with a designated target and in a dual-task paradigm using time-reversed maskers. Thus, the competing voices were not equally important. Mackersie *et al.* (2001)

used pairs of natural sentences and had the listener repeat as much as possible from both sentences, so this was a dual-target, dual-attention task. The segregation skills were then correlated to a simpler psychoacoustic measure of tone fusion in the same hearing-impaired listeners.

A speech segregation test must be designed such that there is no doubt about which signal is the target and which signal is the masker, which will be difficult in the presence of a single competing talker. Furthermore, the ambiguity and equal importance of the two targets should be reflected in the test.

It is also important to consider the use case for a separation algorithm when applied in hearing aids: Ideally, the hearing aids should assist the user in segregating competing voices without any deliberate or very conscious choice made. He or she should simply be able to focus attention on one or few selected targets without instructing or directing the hearing aid. So, a separation algorithm should be designed to accommodate this and be tested in a relevant scenario to evaluate the segregation benefit. If the listening conditions become too difficult, the user may be forced to make a choice that guides the algorithm towards a single target, while suppressing the remaining voices. This will be at the cost of general awareness in the given conversation. Such choice of target could be signaled to the hearing aid from, e.g., a smartphone used as remote control or in the future via EEG or other means of decoding the intention of the listener (O'Sullivan *et al.*, 2017; Perron, 2017).

Similar to Mackersie *et al.* (2001), the competing voices test proposed in this study has two targets that are all equally important to follow, and in the simplified case no masker.

Given the previous work, the present paper has several novel contributions.

- A low-latency (8 ms) deep neural network is used for the separation, as would be needed for a real-time implementation in a hearing aid with an acceptable delay.
- Little training data are used (~3 min recording per speaker), making it realistic for fast training in everyday use.
- Each network instance is trained on two particular voices to optimize subsequent separation of those two voices.
- The aim of the separation is to present the two separated outputs to the two ears to help the hearing-impaired listener focus on one at a time simply by shifting attention.
- The evaluation of the benefit is done using typical, clinically applied speech tests on the target group, the hearing-impaired listeners.

Considering the previous, the main research questions for this study were as follows:

(1) Do hearing-impaired listeners benefit from two signals separated from one mixture and then presented one per ear (dichotic, dual targets)?
(2) Do hearing-impaired listeners benefit from one signal separated from a mixture and presented to both ears (diotic, single target)?
(3) Does the separation benefit depend on the gender mix, e.g., same vs different gender?

II. METHODS

This paper is based on three sub-projects, which will be described in the following.

- The low-latency speech separation algorithm using deep neural networks.
- The speech test method using competing voices in the form of pairs of Danish HINT sentences.
- The evaluation of the proposed speech separation algorithm using the new competing voices test.

A. Low-latency speech separation algorithm using deep neural networks

The speech separation approach used in this study utilizes a deep neural network, which maps spectral features derived from the mixture signal consisting of both talkers to time-frequency masks corresponding to a target talker. Short term Fourier coefficients (STFT) are used as spectral input features and a binary mask/soft ratio mask is used as DNN target output. The present implementation is documented in detail in Naithani *et al.* (2017).

For an acoustic mixture $y(t)$ consisting of sources $s_1(t)$ and $s_2(t)$, the ideal binary mask (Wang and Brown, 2006) corresponding to source $s_1(t)$ can be defined as

$$M_1(t,f) = \begin{cases} 1 & \text{if } |S_1(t,f)| \geq |S_2(t,f)|, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Similarly, the soft ratio mask (Huang *et al.*, 2015) corresponding to source $s_1(t)$ is

$$M_1(t,f) = \frac{|S_1(t,f)|}{|S_1(t,f)| + |S_2(t,f)|}, \quad (2)$$

where $S_1(t,f)$ and $S_1(t,f)$ are STFT spectra corresponding to sources $s_1(t)$ and $s_2(t)$, respectively; time and frequency indices are denoted by, respectively, t and f. Please note that this definition of ratio mask is different from what is referred to as the ideal ratio mask (IRM) (Srinivasan *et al.*, 2006) in literature as it utilizes STFT magnitudes rather than squared STFT magnitudes used in IRM. The DNN is trained in a supervised manner to yield output, $M_1^{est}(t,f)$, the estimated mask corresponding to source $s_1(t)$. The mask corresponding to source $s_2(t)$ is

$$M_2{}^{est}(t,f) = 1 - M_1{}^{est}(t,f). \quad (3)$$

These estimated masks when multiplied with the complex spectra of the mixture signal, $Y(t,f)$, yield the complex STFT spectra of the estimated sources. For first source, it can be expressed as

$$S_1^{est}(t,f) = M_1^{est} \circ Y(t,f), \quad (4)$$

where $\circ$ denotes elementwise multiplication. The other source is extracted similarly. The time domain signals are then recovered via inverse discrete Fourier transform (IDFT) and overlap-add processing. Figure 1 depicts the whole framework for estimation of separated target signal $s_1^{est}(t)$.

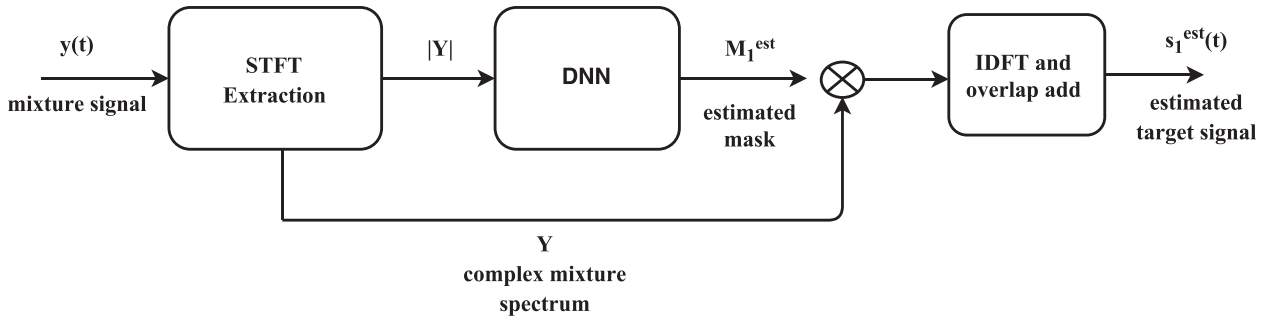

FIG. 1. Illustration of the DNN separation algorithm.

Three types of DNN topologies were utilized in this study: (a) feedforward deep neural network (FDNN), (b) recurrent neural network with long short-term memory units (LSTM), and (c) convolutional recurrent neural network (CRNN). Note that all these topologies were tailored for low algorithmic latency operation, by predicting the mask for only a very short frame at a time, using information from greater previous context. For FDNN, a context of N previous frames were stacked as input when predicting the mask for the current frame in order to provide the past temporal context to the network. In the present study, $N = 4$. No such frame stacking was used for CRNN/LSTM, which unlike FDNN have the inherent capability of modeling the previous temporal context due to the presence of recurrent layers by using the internal state of the network at time *t-1* as its input at time *t*. A detailed description of FDNN and CRNN topologies used in this work can be found from Naithani *et al.* (2016) and Naithani *et al.* (2017), respectively.

Training a DNN implies minimizing an objective function for which the mean square error between the target and estimated masks, i.e., $\sum_{t,f}(M_1(t,f)) - M^{est}{}_1(t,f))^2$, was used. The hyperparameters for the three DNN variants used in Naithani *et al.* (2017) were also used here and are listed in Tables I and II. This parameter selection was carried out using a grid search in order to select the best performing version of each DNN variant. The metric of selection was performance on a validation set in terms of objective metrics of separation, e.g., source-to-distortion ratios (SDR) (Vincent *et al.*, 2006). For the CRNN architecture, a max-pooling operation (Boureau *et al.*, 2010) was performed after each convolutional layer only along the frequency axis in order to preserve the temporal information for the recurrent layers. The rectified linear unit (ReLU) activation (Goodfellow *et al.*, 2016) was used for convolutional layers in CRNN and sigmoid activation (Goodfellow *et al.*, 2016) was used for hidden layers in FDNN. The former choice is conventionally used in convolutional layers (e.g., in Pertilä and Cakir, 2017), the latter choice was directed by experiments on validation data. Similarly, for LSTM units, the conventional activations [hyperbolic tangent (tanh) activation and sigmoid], as in Hochreiter and Schmidhuber (1997), were used. The output activation for all topologies was sigmoid.

TABLE I. Hyperparameters used for FDNN and LSTM networks.

FDNN			LSTM		
Hidden layers	Hidden neurons	Previous context	Hidden layers	Hidden neurons	Sequence length
4	1024	32 ms	3	512	256 ms

In order to prevent overfitting during DNN training, dropout regularization (Srivastava *et al.*, 2014) and early stopping method were used. A dropout value of 0.4 was used and the training was stopped when no improvement in validation loss was observed for 25 epochs. In order to speed up the training, batch normalization (Ioffe and Szegedy, 2015) was used after each hidden/convolutional layer in FDNN/CRNN networks. For gradient optimization, the Adam algorithm (Kingma and Ba, 2015) was used. Librosa (McFee *et al.*, 2017) and Keras (Chollet, 2016) libraries were used for feature extraction and DNN training, respectively.

The HINT speech material used to train the DNNs is described in Sec. II C below. The audio samples were first downsampled to 16 kHz and STFT features were computed using a window length of 8 ms with 50% overlap. This leads to the low algorithmic latency of 8 ms. Four lists, L10, L11, L12, and L13 (see Sec. II C below for the data description), were used as training set and one list, L9, was used as validation set. This amounts to 80 sentences for training and 20 sentences for validation. The training data was generated by first concatenating in time domain all available training lists corresponding to each talker. STFT features were then computed. To increase the amount of training data, multiple training examples were generated by shifting the complex spectrogram of one of the talker relative to the other and summing them to create a mixture. 50 such offsets were used, each resulting in new set of training data. A detailed description of this augmentation procedure can be found in Naithani *et al.* (2017).

B. Objective performance metrics

The performance of the proposed methods was evaluated using objective metrics of separation and intelligibility. For the former, BSS-EVAL toolbox (Vincent *et al.*, 2006) was used for calculating source to distortion ratio (SDR). SDR is used as the measure of overall separation.

For the objective evaluation of intelligibility, extended short-time objective intelligibility (ESTOI) (Jensen and Taal, 2016) was used: The authors postulated that the more widely reported, short-time objective intelligibility (STOI) measure (Taal *et al.*, 2011), does not perform very well in

TABLE II. Hyperparameters used for CRNN. Note that pooling scheme 1 by 2 refers to max pooling operation only along the frequency axis. Convolutional kernel size of 3 by 3 refers to the size of convolutional kernel along time and frequency axes.

Convolutional layers	Recurrent layers	Recurrent neurons	Convolutional filters	Pooling scheme	Sequence length	Convolutional kernel
3	1	256	256	1 by 2	512 ms	3 by 3

situations where highly modulated interferers, e.g., competing speech signal as in our case, are present. ESTOI has been found to work well in these situations as well as where STOI measure works well according to the authors. Hence, we focus the analysis of objective intelligibility metrics on ESTOI, but also report STOI for completeness.

C. The competing voices test with dual targets and single target

The speech test used in this study—the Competing Voices Test (CVT)—has been developed through a series of experiments to arrive at the present version: Pairs of HINT sentences (Nielsen and Dau, 2011) are presented to the listener who is then required to repeat one of them. A visual text cue presented on a second PC monitor indicates to the listener which one of the two sentences to repeat; the cue is either the first or the last word in the target sentence.

We chose to base the new test on existing and proven available Danish speech material, e.g., the Danish HINT, see Nielsen and Dau (2011) and Nilsson *et al.* (1994). The Danish HINT uses natural everyday sentences each containing five words, spoken by one male talker. The entire corpus consists of 13 lists with 20 sentences each: Lists 1–10 are suitable for test, while lists 11–13 have higher spread due to sentence complexity, special words or other reasons (Nielsen and Dau, 2011) and these three lists are then used for training the listener before the actual test starts. Because the training and validation of the speech separation algorithm uses five lists (Sec. II A), only lists 1–8 are available for the listening test.

In order to have multiple talkers to choose from, the HINT sentences with the existing male talker (M1) were re-recorded using two new male talkers (M2, M3) and three new female talkers (F1, F2, F3), hence six talkers in total. The three male talkers spoke with average fundamental frequency (F0) of 100, 130, and 155 Hz and the three females spoke with 200, 172, and 217 Hz. All talkers have the same RMS levels across all sentences, thus the signal-to-signal ratio for two equally important targets is always 0 dB.

D. Experiment I: Evaluation of the source separation algorithm on hearing impaired listeners

1. Test setup

The test was conducted in a soundproof listening booth with headphone presentation via Sennheiser HDA200 audiometry headphones. The test software was a MATLAB application written for the purpose: It handled the experimental protocol, played back the stimuli, and stored the listener responses as entered by the test administrator. All sound files for the experiment were created before the test and stored on the hard disk, and the software would then choose the appropriate sound files, and before playback add hearing loss compensation: Linear gain was prescribed from the individual audiogram according to the NAL-RP gain rationale (Dillon, 2012) and applied via a 256-tap finite impulse response (FIR) filter.

2. Listeners

Fifteen HI listeners were recruited from the Eriksholm test person pool to have moderate-to-severe sloping sensorineural hearing losses, e.g., somewhat flatter than the typical age-induced loss with more low-frequency loss. It was hypothesized that this group would have the largest benefit from the separation algorithm due to a more than mild hearing loss in the low frequencies. Ages were from 47 to 83 years with an average of 73 years. Eight women and seven men participated. The average and range of the hearing losses across both ears on the 15 listeners is shown in Fig. 2.

Apart from the prescribed gain, the overall sound level could be adjusted by the test administrator in dialog with the listener during the training phase to have a comfortable level.

All listeners spoke Danish as their first language. The study was approved by the Research Ethics Committees of the Capital Region of Denmark, and the subjects had signed an informed consent form and were free to withdraw from the experiment at any time. The subjects were not paid for their

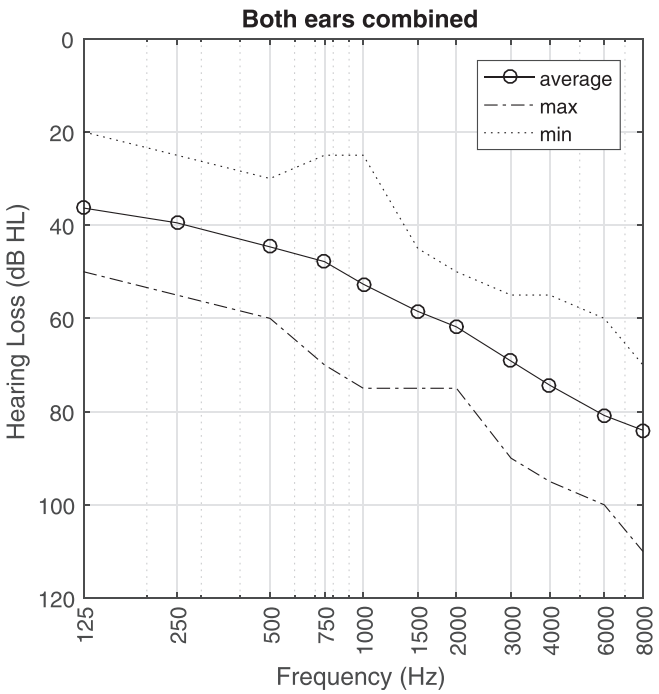

FIG. 2. Summary of 15 listeners' audiograms. Left and right ears are combined.

participation, but they were reimbursed for their travel expenses.

3. Processing modes

With respect to the processing of the stimuli, the following modes were used.

- SUM: The unprocessed sentence mixture.
- SEPARATE: Each unprocessed talker was presented separately. This was the ideal condition corresponding to perfect separation.
- FDNN/LSTN/CRNN: Each DNN separated output was presented separately. It was labeled according to the corresponding DNN used for separation (see Sec. II A for details): Feedforward neural network (FDNN), long-term-short-term neural network (LSTM), and convolutional recurrent neural network (CRNN).

4. Mask mode

The separation mask applied for DNN separation was either binary mask (BM) or ratio mask (RM). For binary mask, the gain applied to each time-frequency cell in the mixture is either zero or 1 to one output and the inverse to the other output. For ratio mask, the gain is a continuous value between 0 and 1 for one output and 1 minus gain for the other output. See Sec. II A for further details. In either case the sum of the two masks is unity at every point, so overall energy of the mixture is preserved. In the case of SEPARATE outputs, the mask was labeled no mask (NM) in the experimental design.

5. Test modes

The listening test presented two sentences from two different HINT lists using two different talkers from the total six talkers. The listener was required to repeat one of the sentences based on a visual cue. The cue was the first or the last word of the target sentence, presented on a screen before or after the sentence pair playback, respectively. In the present study, these two distinct test modes were used:

(1) The CVT test used both sentences as potential targets, thus using both outputs from the DNN or ideal separation. The target word was presented after the playback, requiring equal attention to the two sentences, i.e., a "dual attention" task.
(2) The Target-Masker test (TM) used only the target sentence from the DNN separated (including artefacts) and the SEPARATE conditions and target plus masker (both sentences) in the SUM presentation. The target was indicated to the listener before playback, and the attention was on that target—so this was a "single attention" task. This test mode was similar to existing publications on DNN separation, e.g., Healy *et al*. (2017).

6. Talker pairs and gender mix

The following six talker pairs were used in the test: "M1F1," "M2F3," "M1M2," "M2M3," "F1F2," and "F2F3." This corresponds to two pairs for each of the three gender combination, male-female (MF), male-male (MM), and female-female (FF). Each sentence pair in the listening test was picked from a randomly permutated list of the six talker pairs. Subsequently, the two different talkers were randomly assigned to the left and right sides. During run-time of the listening test, the right or left sentence was randomly chosen as target. The shifting of talkers and positions is more challenging than real-life communication, but was nevertheless chosen here to keep the predictability as low as possible.

7. Experimental design

The experiment thus used the following experimental factors.

- Test mode: CVT, TM (two levels).
- Processing mode: SUM, FDNN, LSTM, CRNN, SEPARATE (five levels).
- Mask mode: BM, RM (two levels). This is nested under processing mode, as the masks are applied for the three DNN processing modes.
- Gender mix: MF, MM, FF (three levels). This was mixed within trials as described above and thus neither generating more trials, nor part of the balancing scheme described below.

Thus, all combinations of Test mode, Processing mode, and Mask mode were covered in the fully factorial design of $2 \times 5 \times 2 = 20$ conditions/trials, and each listener thus heard 20 trials, each containing 20 sentence pairs. The order of conditions was as follows: The two test modes were deliberately ordered, so that all Target-Masker trials always preceded the Competing Voices trials, because they were considered easier and thus good preparation for the harder CVT task. Within each of those two test modes, the remaining 10 conditions were balanced across listeners in a Latin-square fashion to counterbalance any order effects.

8. Procedure

Each trial used the following steps per sentence pair:

- For Target-Masker test mode: A pre-cue (first word of target sentence) is shown on a monitor.
- One or two sentences are played simultaneously depending on test and processing modes.
- For Competing Voices test mode: A post-cue (last word of target sentence) is shown on a monitor.
- The listener repeats (speaks as much of the target sentence as possible) the target sentence according to the cue shown.

All responses were scored per word, i.e., the test administrator marked each of the five words as correct or incorrect according to the rules given by Nielsen and Dau (2011). The word score was then calculated as the sum of the correct words in the last four words in the case of a pre-cue and the correct words in the first four words in the case of a post-cue.

E. Experiment II: Normal-hearing reference data based on ideal separation

Even though the DNN separation algorithm is designed for use with hearing-impaired listeners, it is relevant to know

the expected performance for normal-hearing listeners for two reasons: It can serve as normative data for comparing hearing-impaired results against and for comparing to published normal-hearing results.

For comparing the present results to normal hearing results, an earlier experiment on the competing voices test itself using similar spatial contrasts has been included here (Bramsløw *et al.*, 2015). Due to its explorative nature, only four normal hearing listeners were included, but this was nevertheless enough to obtain statistically valid results. Since it is additional material to the main experiment I, it is here labeled experiment II.

1. Test setup and listeners

The test setup was completely identical to experiment I, however here was no hearing loss compensation applied because of the normal-hearing group. Four normal-hearing listeners from internal staff were included, all with audiometric thresholds <20 dB hearing level. The age span was 28–50 years.

2. Test conditions and procedure

In this experiment, there was no separation algorithm included, but the two reference (spatial) conditions from experiment I were included: SUM (diotic presentation of the two-talker mixture) and SEPARATE (dichotic presentation, one talker per ear). These are the extremes and a separation algorithm should produce results within this range.

The competing voices test paradigm (with post-cue) was identical to that of experiment I, but the target-masker test (with pre-cue) was slightly different, with the masker always present in the opposite ear in the SEPARATE spatial condition.

At the time of this add-on experiment II, only two talkers had been recorded, namely, M1 and F1, so only a Male-Female gender mix was possible. However, the target talker was included as an experimental factor: male/female. The target talker was indicated on a screen as Male or Female.

The procedure was identical to that of Experiment I, but since the target was indicated as Male/Female, all five words in the HINT sentences were available for scoring.

III. RESULTS

A. Experiment I: Hearing-impaired listeners and source separation

1. Analysis method and overview

Upon completion of the test session for all fifteen listeners, all data were collected and analyzed. The word scores in percent were calculated from 0 to 4 correct words, excluding the first or last cue word from the five-word HINT sentence, e.g., $100 * N_{correct}/4$. The 0%–100% word score was then transformed into rationalized arcsine units (rau) as proposed by Studebaker (1985). The rau transformation bends the ends of the psychometric function approximately below 10% and above 90% to make the properties linear and thus better suited for analysis of variance (ANOVA), which assumes normally distributed data. The resulting rau scores range from −18 to 118 rather than 0 to 100.

Now, the rau results were inspected for outliers by grouping either by test persons or processing mode. The motivation was to consider removal of outlying test persons. The outlier range was defined as the 25% and 75% percentiles extended by 1.5 times the 25%–75% distance to either side. Five of the total 900 rau values data points (20 conditions * 3 gender mix = 60 per test person and 15 test persons) fell below the lower outlier limit, but since they were associated with four different test persons, no test persons and thus no data points were removed.

A mixed-model nested factorial ANOVA was run on the rau-transformed word score data with Test Person as a random factor. There were significant main effects of Test mode [$F(1,14) = 119.49$, $p < 0.001$], Processing mode [$F(4,56) = 58.66$, $p < 0.001$], Mask mode (nested under Processing mode) [$F(3,559) = 3.08$, $p < 0.03$], Gender mix [$F(2,28) = 12.87$, $p < 0.0002$], and Test Person [$F(14, 19.84) = 4.67$, $p < 0.001$]. The statistically significant Test Person effect reflects different basic speech recognition skills across listeners in a hearing-impaired group as is often the case in speech tests on such a group, due to both different supra-threshold hearing losses (e.g., Summers *et al.*, 2013) and spread in cognitive function (Lunner, 2003).

Apart from interactions with test person (individual differences), there was one significant interaction: Test mode * Processing mode ($p < 0.001$), meaning that the DNN effect depends on the test mode (Target-Masker vs Competing Voices). The remaining second-order interactions were not significant, neither was the third-order interaction Test mode * Processing mode * Gender mix, so the Test mode * Processing mode interaction did not differ across Gender mix.

The significant interactions with Test Person (TP) were TP * test mode and TP * Test mode * Processing mode. The former interaction indicates different basic performance when comparing the two test modes CVT and TM across listeners, reflecting that the difference in cognitive load between the two tests is more taxing for some listeners than others. The latter interaction furthermore indicates that the effect of processing provides different benefits for different listeners in the two test modes CVT and TM.

In the following, the significant main effects and interaction effects will be presented: All mean rau values from the ANOVA were inverse-transformed to present them as % word scores.

2. Effect of DNN separation

The average main effect of the five processing modes, including the three DNN modes is shown in Fig. 3. *Post hoc* tests showed that SUM is lower than all other conditions (Tukey HSD: $p < 0.001$), SEPARATE is higher than all other conditions (Tukey HSD: $p < 0.001$), and the three DNN modes FDNN, LSTM, and CRNN are not different from one another.

On average, the scores go from 55% (SUM) to 82% (all DNN modes) to 92% (SEPARATE), showing that the DNN separation is substantially higher than SUM ($p < 0.001$) but

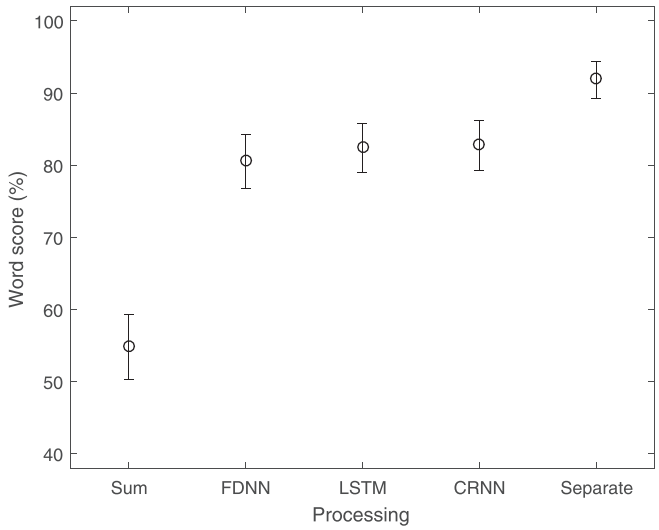

FIG. 3. Word recognition scores for each processing mode, including the three DNN modes. Vertical bars denote 95% confidence intervals.

also inferior to SEPARATE ($p < 0.001$). There is thus potential for improving the DNN separation even further.

The effect of processing mode was different in the two test modes Target-Masker and Competing Voices as seen by the significant interaction Test mode $*$ Processing mode [$F(4,56) = 21.86$, $p < 0.001$]. The corresponding interaction plot is shown in Fig. 4.

The detailed *post hoc* analysis of processing modes within test modes shows the same pattern as the main effect of processing, but for both Test Modes: SUM < "any DNN" < Separated (Tukey HSD: $p < 0.003$). This confirms the observation from the main effect that DNN processing provides a significant benefit over SUM, but not quite as much as the ideal SEPARATE condition. Relative to SUM, the benefits from DNN separation are smaller for the Competing Voices Test than for the Target-Masker test. For the more difficult Competing Voices Test, the scores increase from 52% to an average 64% for the three DNN modes, which is a good benefit in this dual-target situation and 76% in the SEPARATE mode. The benefit for the best DNN is for the LSTM, scoring 65.3% compared to 51.8% for SUM, the rounded word score benefit in this DNN mode is 13%-point. For the Target-Masker test, the scores increase from 58% to approximately 95% in the three DNN modes and 100% in the SEPARATE mode. The DNN benefit is thus roughly 37%-point.

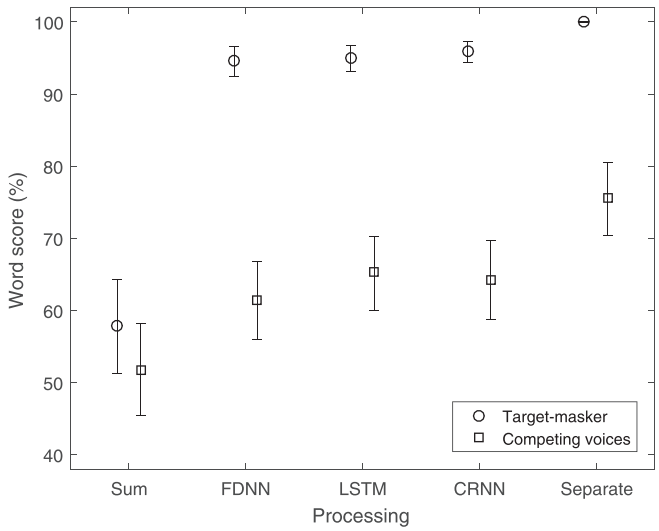

FIG. 4. Word recognition scores as interaction effect of processing mode and test mode, including the three DNN modes. Vertical bars denote 95% confidence intervals. See text for details.

It can furthermore be seen that the SUM condition scores 52%–58% (the difference is not significant) in the two test modes: As the SUM processing is the same in both cases, this means that the pre-cue used in the Target-Masker test does not provide a benefit over the post-cue used in the Competing Voices Test.

3. Effect of binary mask vs ratio mask

It has been an ongoing question in speech separation research whether binary masks or ratio masks provide the higher speech intelligibility scores. It has been claimed that the binary masks provide the better speech intelligibility at the cost of lower sound quality and likewise that the ratio mask provides higher sound quality, as indicated by objective metrics (Wang *et al.*, 2014).

In the present experiment, the mask was nested in the processing mode, because different masks belong to different processing modes: IBM/IRM are specific for DNN processing and "No mask" is used for SUM/SEPARATE, therefore it has been labeled NM for these two processing modes. The effect of Mask mode was significant, but *post hoc* analysis showed that this is driven by the two NM values that are confounded with processing modes SUM and SEPARATE. Inspection of the mask modes BM and RM used with the DNN showed a not significant *post hoc* test and therefore the two mask modes of interest, binary mask vs ratio mask, are not significantly different.

Thus, the choice of mask for DNN separation may be dictated by other requirements than speech intelligibility, e.g., sound quality (not tested here), implementation complexity and cost.

4. Effect of test mode

The test mode effect is simply the difference between the two different test modes Target-Masker with pre-cue and Competing Voices Test with post-cue, including the important difference of presenting only one processed output in the Target-Masker test mode and both processed outputs in the Competing Voices test. Only the SUM processing mode is the same in the two test modes.

The average score in the Target-Masker test is 92% and the score in the Competing Voices Test is 64%, so there is a large difference of 28%-points. This effect has three potential explanations: (1) due to the much simpler task of knowing which talker to attend to, (2) the effect of presenting only one target for the separated conditions in Target-Masker test mode, and (3) finally a designed order effect in the design, because the easier Target-Masker test always preceded the more difficult Competing Voices test. This order effect might increase the CVT scores due to learning effects, thus

reducing the contrast, but likewise decrease the contrast if fatigue was an issue.

5. Effect of gender mix

The effect of gender mix is also relevant to study: if the DNN separation is more difficult for same-gender combinations, this is a weakness in real life applications, as previously found (Isik *et al.*, 2016). In a recent study by Healy *et al.* (2017), only the male-female combination was investigated, and this combination is assumed to be easier for separation, because of the difference in fundamental frequency.

Regarding gender effects, the only significant effect is the main effect of gender, so there are no significant interactions with other experiment factors, specifically, there is no interaction with the processing modes SUM, SEPARATE, etc. In other words, the gender effect is independent of the applied processing. Therefore, only the main effect is shown in Fig. 5.

The average word score is 82% for Male-Female and Male-Male, and significantly lower ($p < 10^{-4}$) for the Female-female combination at 75%. This finding indicates that for the present speech recordings, the female-female combinations are generally slightly more difficult to segregate for the listeners than the two other gender combinations. This finding is specific for the present six talkers and it cannot be generalized to any speech material. Likewise, the unexpected similar results for male-male and male-female may be a result of the quality of the particular talkers used here.

6. Individual effects

As mentioned in Sec. III A 1, the general performance in word score is different across listeners: This reflects different basic speech recognition skills across listeners in a hearing-impaired group as is often the case in speech tests on such a group, due to both different supra-threshold hearing loss and spread in cognitive function (Lunner, 2003). The interaction TP $*$ Processing Mode did not meet the $p < 0.05$ significance criterion [$F(56,44.38) = 1.58$, $p = 0.058$], meaning that overall, there is not a different effect of processing/presentation across the different listeners. This is shown in Fig. 6. However, the spread across listeners is larger for the SUM condition than the SEPARATE condition; this can partly be explained by the ceiling effect. The performance in the three DNN modes is generally close to the performance in the ideal SEPARATE condition. Moreover, for some of the low performers in SUM, a Tukey HSD *post hoc* tests show a significant benefit from DNN separation ($p < 0.05$). In other words, that the poorer performance a listener has in in the unprocessed SUM mode, the more benefit from the DNN separation can be expected.

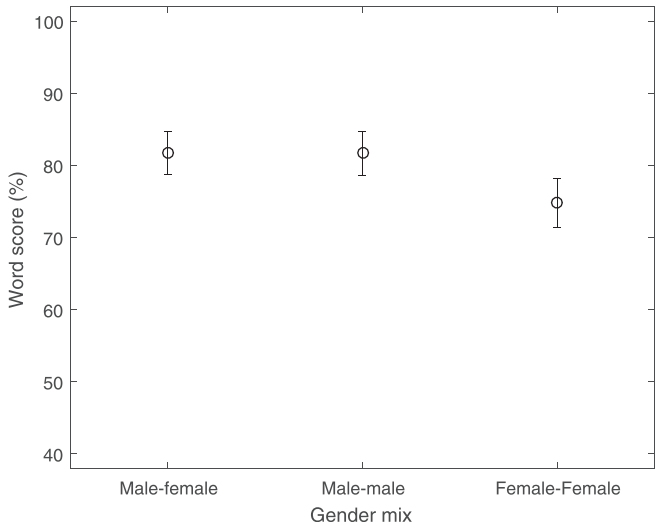

FIG. 5. Word recognition scores for each of the three gender combinations. Vertical bars denote 95% confidence intervals.

7. Objective metrics calculation

As stated in Sec. II, the objective metrics source-distortion ratio (SDR), STOI, and ESTOI were calculated for the audio files presented in the listening test. They were averaged across the experimental conditions from the listening test and are reported in Table III.

For the processing mode, the listening test showed a benefit from DNN separation over SUM. The LSTM score was slightly higher than the two other modes, however the differences between the three DNN modes were not statistically significant (Fig. 3). The objective metrics show a slightly higher SDR for the LSTM than for the two other modes, whereas ESTOI and STOI are almost identical across the three modes. The theoretical metrics for SEPARATE are SDR > 20 dB and STOI/ESTOI $= 1.0$, so the processing can still be improved based on objective scores alone. In the listening test, no effect of binary mask vs ratio mask (mask mode) was found, unlike the objective scores that are consistently higher for the ratio mask. This improvement did apparently not translate into a user benefit. For the gender mix, there was a small and statistically significant decline in the listening test for the FF pairs, and this decline can also be observed for both the SDR and the STOI/ESTOI scores.

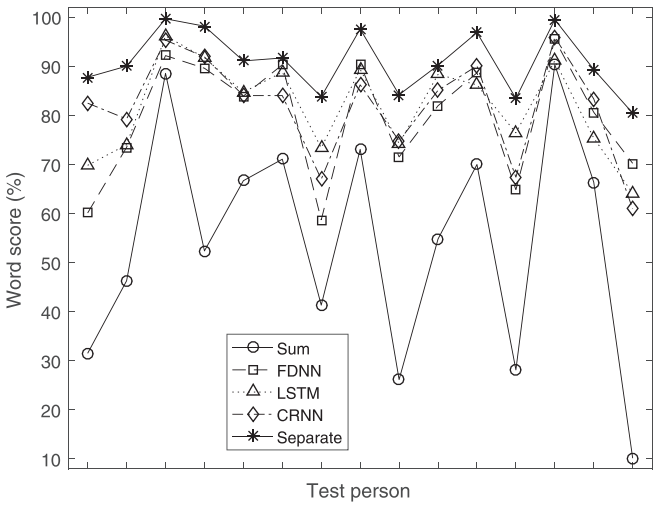

FIG. 6. The average word scores per listener (TP). The interaction is not statistically significant overall, but *post hoc* tests show a statistically significant benefit from the DNN separation for the listeners with the lowest SUM performance.

TABLE III. Objective metrics for the conditions used in the listening test in experiment I. The scores for SEPARATE are not measured but theoretical ideal values.

		SDR(dB)	ESTOI	STOI
Proc mode	SUM	0.25	0.55	0.71
	FDNN	5.20	0.70	0.83
	LSTM	5.29	0.71	0.83
	CRNN	5.24	0.71	0.83
	SEPARATE	>20	1.00	1.00
Mask mode	BM	4.83	0.67	0.80
	RM	5.66	0.75	0.86
Gender mix	MF	6.04	0.74	0.85
	MM	5.22	0.72	0.85
	FF	4.48	0.67	0.79

Although ESTOI was designed for fluctuating maskers, it did not show different trends than STOI with the competing voices used here, only the ESTOI scores were generally lower.

B. Experiment II: Normal-hearing listeners and ideal separation

1. Analysis method and overview

The data from the four normal-hearing listeners were rau-transformed (Studebaker, 1985) prior to ANOVA. A mixed-model ANOVA was performed, with TP as a random factor. The following main effects were significant: TP [$F(3.6, 243) = 7.9$, $p = 0.044$], Processing [$F(6, 271) = 7.7$, $p = 0.02$] and Test mode [$F(6, 143) = 14.5$, $p = 0.005$]. Furthermore, some significant effects were found: TP $*$ Processing [$F(112, 99.7)$, $p = 0.02$] and Processing $*$ Test Mode [$F(245,112) = 2.46$, $p = 0.05$].

2. Effect of spatial separation and test mode

The latter interaction thus summarizes all effects of Processing and Test mode and is shown in Fig. 7.

For the target-masker test (cued before sentence pair presentation), the scores are 97% and 99% for the Sum and Separate processing modes, very close to ceiling. For the competing voices test mode (cued after sentence pair presentation), the scores are 90% and 98% for the Sum and Separate modes. The difference is significant according to a Tukey HSD *post hoc* test ($p < 0.001$).

As stated previously, experiment II preceded experiment I, and we did thus not test the DNN separation on normal-hearing listeners. However, according to Fig. 7, the outcome from testing competing voices (post cue) of applying DNN separation can be predicted to lie between 90% and 98%, very close to ceiling, and in any case a small benefit in a range where normal-hearing listeners perform quite well.

IV. DISCUSSION

A. Benefit of DNN separation

In the competing voices scenario, the user is maintaining attention on both targets, and has both available, for voluntary shift of attention. The aim of the proposed algorithm is to facilitate this type of communication scenario by presenting the two separated talkers dichotically. For the hearing-impaired listeners in experiment I, this benefit is observed in Fig. 4 (square symbols) as a statistically significant effect of approximately 13%-point, which is worthwhile for a hearing-impaired individual who needs help in these challenging competing voice scenarios. The benefit is achievable because of a rather low average base performance in the SUM condition around 55% with a large spread across listeners as shown in Fig. 6.

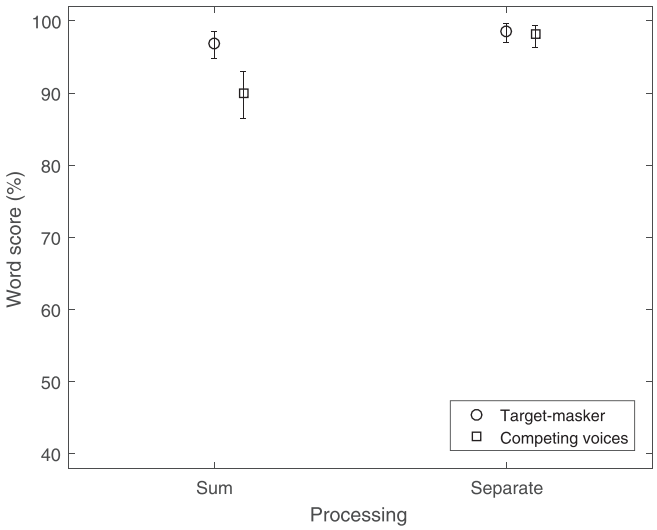

FIG. 7. The combined effects of processing mode and test mode for four normal-hearing listeners. The plot uses the same y-axis limits as the previous plots for the hearing-impaired group, e.g., Fig. 4 showing the same results for the hearing-impaired group.

When there is a designated target (Fig. 4, circle symbols), the effect of DNN separation on word scores is very high for the hearing-impaired group. This scenario is relevant when the user has chosen what the preferred target is, by, e.g., indicating on a remote control or indicating the target indirectly by events that can be detected such as head or eye movements (Kidd *et al.*, 2013). In this case, the benefits from speech separation were substantial, around 37%-point improved word score. In comparison, Healy *et al.* (2013) reported scores going from near zero to values above 70% for speech shaped noise and babble while Healy *et al.* (2017) obtained an average of 42.5%-point for SNR at −3 dB and 59%-point benefit for SNR at −9 dB for a male target talker in the presence of a female competing talker. In the present study, lower scores at the average SNR of 0 dB were found, and the benefit was independent of the gender mix. The separation can thus be beneficial for continued dual target attention, or for helping in choosing a target and thus obtaining the higher single-target benefit. In a future scenario, this could be done via cognitive control of a hearing aid (Perron, 2017).

For the fixed SNR of 0 dB as used in the present use case, a segregation benefit in normal-hearing listeners from DNN separation could also be expected, as shown by, e.g., Healy *et al.* (2017) at slightly more difficult SNRs: At −6 dB, they reported 6.1%-points benefit from

approximately 88% to 94% in a test scenario equivalent to the Target-Masker test applied in the present study. This benefit would likely disappear at 0 dB SNR because the unprocessed sum would have scores very close to 100% ceiling. While the present study did not test the DNN algorithms on NH listeners, experiment II for SUM/SEPARATE was conducted and is summarized in Fig. 7: For the target-masker, a SUM score of 97% leaves little room for benefit and for competing voices, the SUM score is 90% and SEPARATE is 98%, so a small benefit could be expected, but the use case is not very relevant so close to ceiling performance.

B. Tolerance to gender mix

Unlike earlier published results (Isik *et al.*, 2016), the separation algorithms used here did not perform differently for different combinations of male and female talkers within the employed set of three males and three females. This is a positive outcome from the users' and developers' perspective, because this gives the same benefit in all talker combinations.

The average scores across gender (Fig. 5) combinations do show a slightly lower word recognition score for the female-female combination compared to the two other combinations male-female and male-male. With the particular six talkers used in this study, there is thus no different-gender advantage, unlike what was found by Brungart (2001).

C. Limitations

The current experiment investigated the benefit of separating two mixed, known voices. The circumstances were ideal in the sense that there was no background noise neither during training nor during test of the separation algorithm. We did not test the effects of unknown talker or unknown background noises, but since the generalization to new talkers is not perfect, some reduction of the benefit can be expected (Kolbæk *et al.*, 2017; Kumar and Florencio, 2016).

In other aspects, the present experiment imposed restrictions that may be relaxed in future work: The hearing-aid critical latency restriction of 8 ms may be expanded to improve separation performance, as found by Barker *et al.* (2015), possibly at a cost in terms of reduced sound quality (Bramsløw, 2010; Stone *et al.*, 2008). Likewise, the restricted training material of approximately 3 min is far from ideal for a large DNN and a larger speech corpus for training could lead to better separation performance (Kolbæk *et al.*, 2017). However, at the same time, the small data requirement increases the applicability in real world situations as only a brief speech sample is required for learning a voice and use it in separation.

The special use of the HINT material using only 8 lists with 20 sentences each for the evaluation on hearing-impaired listeners will inevitably lead to some learning by the listeners. We have compensated for this by using a balanced experimental design, thus balancing out the learning effects. However, a larger speech corpus for testing could ease this problem, or even using alternative outcome measures such as pupillometry or EEG as a proxy for listening effort (Koelewijn *et al.*, 2014).

In a recent study, Ohlenforst *et al.* (2017) used pupil dilation measurements to assess listening effort as function of signal-to-noise ratio (SNR) when listening to speech masked by either stationary noise or a single competing voice. In both normal-hearing and hearing-impaired listeners, the pupil dilation was largest for sentence recognition scores around 50% and then decreasing when a higher SNR was used, leading to higher word scores. This indicates that the listening effort is diminished when the speech recognition is improved.

D. Objective measurements of separation

Generally, the objective scores SDR, ESTOI, and STOI listed in Table III agree well with the results from the listening test in experiment I: The objective scores for DNN separation are higher than for the unprocessed SUM and lower than the ideal SEPARATE, and furthermore, the scores are the same for the three modes. The two mask types, Binary Mask and Ratio Mask, were not different in the listening test, but in the objective scores, Ratio Mask is ∼0.8 dB higher for SDR and 0.08/0.06 for ESTOI/STOI. The three gender mixes show the same pattern in the objective scores for ESTOI/STOI: Male-female and male-male show practically the same scores and female-female are slightly below. For the SDR, there is furthermore a difference between male-female and male-male, which was not found in the listening test.

The best scores for DNN separation are for the ratio mask: SDR = 5.66 dB and, ESTOI = 0.75 and STOI = 0.86. The average SNR across sentence pairs is 0 dB in the competing voices test. In a similar study, Healy *et al.* (2017) measured STOI for DNN separation of one talker from a male-female mixture of sentence pairs using ratio mask, and obtained slightly higher STOI scores of 0.91 at −3 dB SNR. The augmented amount of training data is similar in the present study and in Healy *et al.* (2017), on the order of 9000 s, however, the present study was trained using only 80 sentences, compared to 600 sentences in Healy *et al.* (2017).

For a single-talker masker as used in the present study, the more appropriate ESTOI (Jensen and Taal, 2016) did not show different trends than STOI (Taal *et al.*, 2011), but had generally lower scores.

E. Perspectives

In the present study, we have focused on two distinctly different use cases of the two-talker scenario.

- Competing voices: The user is attending to both voices and attempting to focus on both talkers. This requires no user input to the hearing system—the two separated outputs are presented separately to the two ears. The segregation enhancement and improved speech intelligibility is beneficial as such and available to the user without any further action—enabling fast voluntary attention shifting.
- Target-Masker: The user makes a choice of which of the two voices to attend to and selects this as the target over

the masking other voice. The user must somehow inform the system which of the two outputs is in focus, by, e.g., switching on a smartphone app or using eye gaze (Kidd *et al.,* 2013), hereby leaving out the possibly relevant information from the other voice. The DNN separation of one talker from a mixture of multiple talkers and automatic selection via EEG has also been suggested by O'Sullivan *et al.* (2017), however, this required intracranial EEG recordings in this first proof-of-concept.

The present study shows a statistically significant benefit from the DNN separation in both cases, roughly 13%-point in the CV listening and around 37%-points in the TM listening. Both benefits are statistically significant and valuable to the hearing-impaired users, who often find themselves struggling in these situations.

The current efforts should be regarded as a simplified laboratory-based proof of concept, and as such, it was successful. Due to the choice of a well-known speech corpus already used for hearing aid evaluations—HINT—there are also limitations in the study that will not occur in real life, e.g., memorization of the sentences and the limitation in training material to four lists, equivalent of 3 min speech. In the real-world application, more useful training data can be accumulated over time, but also more noisy data than used presently. The implications of these changes must be investigated.

The present implementation of the separation algorithm is based on STFT spectra and as such, there is no audiological knowledge built-in. It could be advantageous to incorporate elements of hearing loss, spread of masking (e.g., Launer and Moore, 2003; Goehring *et al.,* 2016) as in an auditory model into the algorithm.

Finally, it may be worthwhile investigating other outcome measures than word scores, especially indirect measures that use running speech as the main stimulus. One such example would be using pupillometry for assessing the benefit of the separation algorithm (Koelewijn *et al.,* 2014), by measuring pupil dilation as an indicator of listening effort. Benefits in speech recognition scores will generally lower the pupil dilation and hence listening effort (Ohlenforst *et al.,* 2017).

V. SUMMARY AND CONCLUSIONS

In the work presented here, we have tested the effect of a speech separation algorithm based on DNNs and applied it to achieve segregation enhancement for hearing-impaired listeners. The algorithm is targeted towards a hearing-aid application, thus using short latencies (8 ms) and little training data (∼3 min) compared to similar published algorithms. It was evaluated on 15 hearing-impaired listeners using a new competing voices test and showed a benefit in two types of test scenarios: a competing voices scenario with divided attention on two equally important voices (dual target, dual-attention) to demonstrate segregation enhancement and a target-masker scenario with selected focus on one target (single target, singe attention). Both scenarios and benefits are relevant for the end users.

A smaller experiment on four normal-hearing listeners showed performance near ceiling (90%–100% word score) for both mixed and separated competing voices, thus leaving little room for a potential segregation benefit from DNN separation in this group.

In the competing voices (dual-target) scenario the listener is aware of both voices and can attend voluntarily to one or the other. No other intervention or guidance of the algorithm is required. In this rather demanding use case, the separation algorithm provided a benefit of approximately 13%-points. As the access to both voices was improved, this may enable the listener to catch words from a competing conversation and decide to switch attention, or simply tune in conversations in a cocktail party—or during family dinners.

In a more classic target-masker scenario, one voice is identified as target and the other as masker. Thus, the listener must actively indicate to the algorithm which voice is the target and the other voice will be suppressed. The indication can be done via, e.g., a smartphone application. In this case, the separation benefit was approximately 37%-point, and dual attention is not possible.

The listening test results for the hearing-impaired group were very well in line with popular objective metrics, thus confirming the usefulness of these metrics in further optimization of the DNN separation algorithm.

The limitations of the algorithm in terms of robustness to new voices not yet trained, noise in training and test data, and voice variations over time remains to be evaluated. The benefit from more training material than the current 3 min should also be investigated.

ACKNOWLEDGMENTS

The work from Tampere University of Technology was partly funded by Grant No. 15-0653 from the Oticon Foundation. We thank Jette Nissen for help with booking of the listeners and CSC-IT Centre of Science Ltd., Finland, for providing computational resources.

Bach, F. R., and Jordan, M. I. (**2005**). "Blind one-microphone speech separation: A spectral learning approach," Adv. Neural Inf. Process. Syst. **17**, 65–72.

Barker, T., Virtanen, T., and Pontoppidan, N. H. (**2015**). "Low-latency sound-source-separation using non-negative matrix factorisation with coupled analysis and synthesis dictionaries," in *2015 IEEE International Conference on Acoustics and Speech Signal Processes*, IEEE, pp. 241–245.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Boll, S. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. **27**, 113–120.

Boureau, Y.-L., Ponce, J., and LeCun, Y. (**2010**). "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th International Conference on Machine Learning*, pp. 111–118.

Bramsløw, L. (**2010**). "Preferred signal path delay and high-pass cut-off in open fittings," Int. J. Audiol. **49**, 634–644.

Bramsløw, L., Vatti, M., Hietkamp, R. K., and Pontoppidan, N. H. (**2015**). "Binaural speech recognition for normal-hearing and hearing-impaired listeners in a competing voice test," in *Speech Noise 2015*, Copenhagen.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Chandna, P., Miron, M., Janer, J., and Gómez, E. (**2017**). "Monoaural audio source separation using deep convolutional neural networks," in *International Conference on Latent Variable Analysis and Signal Separation* (Springer, Berlin), pp. 258–266.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.
Chollet, F. (**2016**). *Keras*, GitHub, https://github.com/keras-team/keras/releases/tag/1.1.0 (Last viewed June 29, 2018).
Dillon, H. (**2012**). *Hearing Aids*, 2nd ed. (Thieme, New York).
Erdogan, H., Hershey, J. R., Watanabe, S., and Le Roux, J. (**2015**). "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015*, pp. 708–712.
Ezzatian, P., Li, L., Pichora-Fuller, K., and Schneider, B. A. (**2015**). "Delayed stream segregation in older adults," Ear Hear. **36**, 482–484.
Goehring, T., Yang, X., Monaghan, J. J. M., and Bleeck, S. (**2016**). "Speech enhancement for hearing-impaired listeners using deep neural networks with auditory-model based features," in *2016 24th European Signal Processing Conference*, IEEE, pp. 2300–2304.
Goodfellow, I., Bengio, Y., and Courville, A. (**2016**). *Deep Learning* (MIT Press, Cambridge, MA).
Grais, E. M., Sen, M. U., and Erdogan, H. (**2014**). "Deep neural networks for single channel source separation," in *2014 IEEE International Conference on Acoustics and Speech Signal Processing*, IEEE, pp. 3734–3738.
Han, K., and Wang, D. (**2012**). "A classification based approach to speech segregation," J. Acoust. Soc. Am. **132**, 3475–3483.
Hanson, B. A., and Wong, D. Y. (**1984**). "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 195–199.
Healy, E. W., Delfarah, M., Vasko, J. L., Carter, B. L., and Wang, D. (**2017**). "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," J. Acoust. Soc. Am. **141**, 4230–4239.
Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660–1669.
Healy, E. W., Yoho, S. E., Wang, Y., Apoux, F., and Wang, D. (**2014**). "Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **136**, 3325–3336.
Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.
Helfer, K. S., Chevalier, J., and Freyman, R. L. (**2010**). "Aging, spatial cues, and single-versus dual-task performance in competing speech perception," J. Acoust. Soc. Am. **128**, 3625–3633.
Hochreiter, S., and Schmidhuber, J. (**1997**). "Long short-term memory," Neural Comput. **9**, 1735–1780.
Huang, P. Sen, Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (**2015**). "Joint optimization of masks and deep recurrent neural networks for monaural source separation," IEEE/ACM Trans. Speech Lang. Process. **23**, 2136–2147.
Ihlefeld, A., and Shinn-Cunningham, B. (**2008**). "Disentangling the effects of spatial cues on selection and formation of auditory objects," J. Acoust. Soc. Am. **124**, 2224–2235.
Ioffe, S., and Szegedy, C. (**2015**). "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456.
Isik, Y., Roux, J. Le, Chen, Z., Watanabe, S., and Hershey, J. R. (**2016**). "Single-channel multi-speaker separation using deep clustering," in *Proceedings of INTERSPEECH, ISCA*, pp. 545–549.
Jang, G. J., and Lee, T. W. (**2004**). "A maximum likelihood approach to single-channel source separation," J. Mach. Learn. Res. **4**, 1365–1392.
Jensen, J., and Taal, C. H. (**2016**). "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE/ACM Trans. Audio Speech Lang. Process. **24**, 2009–2022.
Kidd, G., Favrot, S., Desloge, J. G., Streeter, T. M., and Mason, C. R. (**2013**). "Design and preliminary testing of a visually guided hearing aid," J. Acoust. Soc. Am. **133**, EL202–EL207.
Kingma, D. P., and Ba, J. L. (**2015**). "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, pp. 1–15.
Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., and Kramer, S. E. (**2014**). "The pupil response is sensitive to divided attention during speech processing," Hear. Res. **312**, 114–120.
Kolbæk, M., Tan, Z.-H., and Jensen, J. (**2017**). "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," IEEE/ACM Trans. Audio Speech Lang. Process. **25**, 153–167.
Kollmeier, B., and Wesselkamp, M. (**1997**). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," J. Acoust. Soc. Am. **102**, 2412–2421.
Kumar, A., and Florencio, D. (**2016**). "Speech enhancement in multiple-noise conditions using deep neural networks," arXiv:1605.02427.
Launer, S., and Moore, B. C. (**2003**). "Use of a loudness model for hearing aid fitting V on-line gain control in a digital hearing aid," Int. J. Audiol. **42**, 262–273.
Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (**2013**). "Speech enhancement based on deep denoising autoencoder," in *Interspeech*, pp. 436–440.
Lunner, T. (**2003**). "Cognitive function in relation to hearing aid use," Int. J. Audiol. **42**, S49–S58.
Luo, Y., and Mesgarani, N. (**2017**). "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," arXiv:1711.00541.
Mackersie, C. L., Prida, T. L., and Stiles, D. (**2001**). "The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss," J. Speech Lang. Hear. Res. **44**, 19–28.
McFee, B., McVicar, M., Nieto, O., Balke, S., Thome, C., Liang, D., Battenberg, E., Moore, J., Bittner, R., Yamamoto, R., Ellis, D., Stoter, F.-R., Repetto, D., Waloschek, S., Carr, C., Kranzler, S., Choi, K., Viktorin, P., Santos, J. F., Holovaty, A., Pimenta, W., and Lee, H. (**2017**). Librosa 0.5.0.
Naithani, G., Barker, T., Parascandolo, G., Bramsløw, L., Pontoppidan, N. H., and Virtanen, T. (**2017**). "Low-latency sound source separation using convolutional recurrent deep neural networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, New Paltz, NY, pp. 1–5.
Naithani, G., Parascandolo, G., Barker, T., Pontoppidan, N. H., Virtanen, T., Parascandolo, G., Bramsløw, L., Pontoppidan, N. H., and Virtanen, T. (**2016**). "Low-latency sound source separation using deep neural networks," in *2016 IEEE Global Conference on Signal and Information Processing*, IEEE, pp. 272–276.
Naylor, J., and Boll, S. (**1987**). "Techniques for suppression of an interfering talker in co-channel speech," in *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 12.
Neher, T., Behrens, T., Kragelund, L., and Petersen, A. S. (**2007**). "Spatial unmasking in aided hearing-impaired listeners and the need for training," in *Proceeding of the International Symposium on Auditory and Audiological Research*, Helsingør, Denmark, pp. 515–522.
Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol. **50**, 202–208.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., and Kramer, S. E. (**2017**). "Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation," Hear. Res. **351**, 68–79.
O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., and Mesgarani, N. (**2017**). "Neural decoding of attentional selection in multi-speaker environments without access to clean sources," J. Neural Eng. **14**, 056001.
Park, S. R., and Lee, J. (**2016**). "A fully convolutional neural network for speech enhancement," arXiv:1609.07132.
Parsons, T. W. (**1976**). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. **60**, 911–918.
Perron, M. (**2017**). "Hearing aids of tomorrow: Cognitive control toward individualized experience," Hear. J. **70**, 22–23.
Pertila, P., and Cakir, E. (**2017**). "Robust direction estimation with convolutional neural networks based steered response power," in *2017 IEEE International Conference on Acoustics and Speech Signal Processing*, IEEE, pp. 6125–6129.
Pontoppidan, N., and Dyrholm, M. (**2003**). "Fast monaural separation of speech," in *23rd International Conference on Signal Processing and Audio Recording Reproduction*, pp. 1–6.

Quatieri, T. F., and Danisewicz, R. G. (**1990**). "An apporoach to co-channel talker interference suppression using a sinusoidal model for speech," IEEE Trans. ASSP **38**, 56–69.
Raj, B., and Smaragdis, P. (**2005**). "Latent variable decomposition of spectrograms for single channel speaker separation," in *IEEE Workshop on Applied Signal Processing to Audio Acoustics*, pp. 17–20.
Roman, N., and Wang, D. (**2006**). "Pitch-based monaural segregation of reverberant speech," J. Acoust. Soc. Am. **120**, 458–469.
Roweis, S. T. (**2001**). "One microphone source separation," Adv. Neural Inf. Process. Syst. **13**, 793–799.
Seltzer, M. L., Raj, B., and Stern, R. M. (**2000**). "Classifier-based mask estimation for missing feature methods of robust speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, Vol. 3, pp. 538–541.
Srinivasan, S., Roman, N., and Wang, D. (**2006**). "Binary and ratio time-frequency masks for robust speech recognition," Speech Commun. **48**, 1486–1501.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (**2014**). "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res. **15**, 1929–1958.
Stone, M. A., Moore, B. C., Meisenbacher, K., and Derleth, R. P. (**2008**). "Tolerable hearing aid delays. V. Estimation of limits for open canal fittings," Ear Hear. **29**, 601–617.
Stubbs, R. J., and Summerfield, Q. (**1990**). "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **87**, 359–372.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Lang. Hear. Res. **28**, 455–462.
Summers, V., Makashay, M. J., Theodoroff, S. M., and Leek, M. R. (**2013**). "Suprathreshold auditory processing and speech perception in noise: Hearing-impaired and normal-hearing listeners," J. Am. Acad. Audiol. **24**, 274–292.
Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.
Tamura, S., and Waibel, A. (**1988**). "Noise reduction using connectionist models," in *ICASSP*, pp. 553–556.
Vincent, E., Gribonval, R., and Fevotte, C. (**2006**). "Performance measurement in blind audio source separation," IEEE Trans. Audio Speech Lang. Process. **14**, 1462–1469.
Virtanen, T. (**2007**). "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," IEEE Trans. Audio Speech Lang. Process. **15**, 1066–1074.
Wagener, K., Josvassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.
Wang, D. (**2008**). "Time-frequency masking for speech separation and its potential for hearing aid design," Trends Amplif. **12**, 332–353.
Wang, D., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE, New York), Vol. 147, pp. 147–185.
Wang, D., and Chen, J. (**2017**). "Supervised speech separation based on deep learning: An overview," arXiv:1708.07524.
Wang, D., and Hu, G. (**2006**). "Unvoiced speech segregation," in *2006 IEEE International Conference on Acoustics and Speed Signal Processing*, IEEE, pp. V-953–V-956.
Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.
Wang, Y. (**2015**). "Supervised speech separation using deep neural networks," Ph.D. thesis, Ohio State University.
Wang, Y., Narayanan, A., and Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE Trans. Acoust. Speech Lang. Process. **22**, 1849–1858.
Wang, Y., and Wang, D. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio Speech Lang. Process. **21**, 1381–1390.
Weninger, F., Hershey, J. R., Le Roux, J., and Schuller, B. (**2014**). "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal Informaton Processing*, IEEE, pp. 577–581.
Williamson, D. S., and Wang, D. L. (**2017**). "Time-frequency masking in the complex domain for speech dereverberation and denoising," IEEE/ACM Trans. Audio Speech Lang. Process. **25**, 1492–1501.
Xie, F., and Van Compernolle, D. (**1994**). "A family of MLP based nonlinear spectral estimators for noise reduction," in *Proceedings of ICASSP'94. IEEE International Conference on Acoustics and Speech Signal Processing*, pp. II/53–II/56.
Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (**2014**). "An experimental study on speech enhancement based on deep neural networks," IEEE Sign. Process. Lett. **21**, 65–68.
Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (**2015**). "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Trans. Audio Speech Lang. Process. **23**, 7–19.