INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Optimization and evaluation of an intelligibility-improving signal processing approach (IISPA) for the Hurricane Challenge 2.0 with FADE

Marc René Schädler

Medical Physics and Cluster of Excellence Hearing4All
marc.r.schaedler@uni-oldenburg.de

Abstract

This contributions describes the "IISPA" submission to the Hurricane Challenge 2.0. The challenge organizers called for submissions of speech signals processed with the aim to improve their intelligibility in adverse listening conditions. They evaluated the submissions with matrix sentence tests in an international listening experiment. An intelligibility-improving signal processing approach (IISPA) inspired from research on speech perception of listeners with impaired hearing was designed. Its parameters were optimized with an objective intelligibility model, the simulation framework for auditory discrimination experiments (FADE). In FADE, a re-purposed automatic speech recognition (ASR) system is employed as a models for human speech recognition performance. The model predicted an improvement in speech recognition threshold (SRT) of approximately 5.0 dB due to the optimized IISPA. The processed speech signals were evaluated in the Hurricane Challenge 2.0. The measured improvements were language-dependent: up to 4.8 dB for the Spanish test, up to 3.8 dB for the German test, and up to 2.1 dB for the English test. The results show on the one hand the potential of using an ASR-based speech recognition model to optimize an intelligibility-improving signal processing scheme, and on the other hand the need for thorough listening experiments.

Index Terms: speech enhancement, intelligibility improvement, automatic speech recognition, speech recognition model

1. Introduction

We rely on speech as our primary communication channel in many daily situations. Sometimes, speech is the only available source of information, e.g., announcements in a mall or an airplane. Then, noise and reverberation can make the correct recognition of important information difficult or even impossible. In this case, highly intelligible speech can be crucial for the successful communication. Intelligibility-enhancing modifications of the speech signal were shown to improve the speech recognition performance in noisy and reverberant listening conditions [1].

The Hurricane Challenge [1] was the first large-scale open evaluation of intelligibility-improving algorithms, where eighteen contributions competed on a common data set to increase speech intelligibility in noise. Several contributions relied on spectral shaping, dynamic range compression, as well as spectral contrast enhancements. Some used speech-intelligibility prediction models, like the speech intelligibility index [2], to optimize the signal processing parameters. Most objective intelligibility models (or metrics) (OIMs) rely on signal-to-noise ratios (SNRs) or correlations between representations of the processed and unprocessed speech and fail to provide accurate predictions for processed (speech) signals [3]. Also, these metrics do not predict the outcome of a speech recognition experiment but an index values which is designed to be highly correlated with the empirical speech recognition performance.

Recently, a simulation framework for auditory discrimination experiments (FADE, [4]) was proposed that employs a repurposed automatic speech recognition (ASR) system to simulate speech recognition tests. An advantage of this approach is, that it predicts the outcome of a speech recognition test, e.g., a speech recognition thresholds (SRT) or a recognition rate for a given SNR. The approach was shown to predict the effect of stationary and fluctuating noise maskers on the outcome of the matrix sentence test in several languages [5]. Further, it was shown to accurately predict the improvement in SRT due to noise reduction schemes, i.e., the speech recognition performance with non-linearly-processed noisy speech signals [6]. With a modification to the feature-extraction stage of the ASR system, it was even shown to accurately predict the individual benefit in SRT in noise due to different hearing-loss compensation strategies including compression amplification [7]. Hence, the approach could be suitable to optimize intelligibility-enhancing signal processing algorithms.

The Hurricane Challenge 2.0 [8] again performed a coordinated international evaluation of the intelligibility of modified speech in Spanish, English, and German language. The participants were provided with unprocessed matrix sentence speech test signals in three languages, impulse responses, corresponding noise signals, and a method to generate noisy speech signals at three positions with different reverberation. The task was to improve the speech recognition performance by processing the unprocessed speech signals prior to adding reverberation and mixing them with the noise signal. The evaluation was performed with over 60 native listeners in each language, where positions, and hence the impulse responses and noises, were slightly different from those provided for the optimization.

In this contribution, the parameters of an intelligibility-improving signal processing approach (IISPA) were optimized for the Hurricane Challenge 2.0 using FADE simulations. The predicted and achieved improvements in the Hurricane Challenge 2.0 are presented and discussed.

2. Methods

2.1. Intelligibility-improving signal processing approach

The IISPA was designed with the aim to expose a few parameters which could be optimized using FADE simulations. A GNU/Octave reference implementation documents the implementation details[1]. Here, the core processing blocks are described and illustrated. For the analysis of the speech signal, a spectro-temporal representation that is widely used in the feature-extraction stages of ASR systems, the logMS was used. An example of a logMS of a portion of an unprocessed speech signal can be observed in the top panel of Figure 1. The spectral resolution was approximately 1 ERB (equivalent rectangu-

[1] https://github.com/m-r-s/fade-hurricane

http://dx.doi.org/10.21437/Interspeech.2020-93

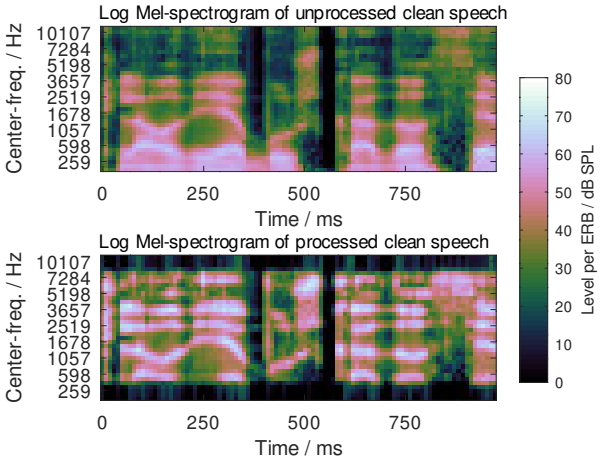

Figure 1: *Log Mel-spectrogram of an unprocessed Spanish speech sample (upper panel) and the corresponding signal processed with IISPA (lower panel), both normalized to 65 dB SPL.*

lar bandwidth), i.e., the bandwidth of an auditory filter in the human auditory system. The analysis window shift was 10 ms (the window length 25 ms) and allowed to use an overlap-add method to modify the corresponding waveform in 20 ms windows. Gains for each time-frequency bin of the logMS were determined and applied as follows:

1) Spectral gains in dB were determined by a polynomial function of degree 2 on a logarithmic frequency scale to the base 2 centered at 2 kHz. The slope, which can be interpreted in dB per octave, and curvature were exposed as optimization parameters. The slope parameter S can weight frequencies below 2 kHz against frequencies above 2 kHz, and the curvature parameter C frequencies close to 2 kHz against frequencies far from 2 kHz.

2) Band-pass characteristic with a lower cut-off frequency of L=500 Hz and an upper cut-off frequency of U=8000 Hz was applied. The gains for bands of the logMS with center frequencies outside this range were set to $-\infty\, dB$. The values of 500 Hz and 8000 Hz for L and U, respectively, were determined in pilot experiments and fixed for all languages. The effect of the band-pass can be observed in the example of processed speech in the lower panel of Figure 1.

3) Dynamic range manipulation of the logMS such that signal dynamics which are relevant for speech recognition were preserved and dynamic which is less relevant was removed. The following description is strongly related to work on a possible compensation strategy for hearing loss which is patented in Germany (DE 10 2017 216 972). The logMS of the unprocessed signal was convolved with seven spectro-temporal smoothing kernels. The effective sizes of the used Hanning-window kernels are listed in Table 1. An illustration the result of the convolutions, i.e., examples of the smoothed versions of the logMS

Table 1: *Spectral and temporal smoothing kernel (Hanning window) sizes used to derive smoothed versions of the logMS of the unprocessed signal, in Mel-bands and ms, respectively.*

Layer	1	2	3	4	5	6	7
Spectral	1	2	4	4	8	16	32
Temporal	10	10	10	30	30	30	30

which are referred to as layers, are depicted in the left row in Figure 2. The element-wise differences between adjacent layers are depicted in the right column. The differences provide a decomposition of the logMS in spectral and temporal modulations.

Here, the smoothing can be interpreted as a spectro-temporal modulation low-pass filtering, and differences between low-passes can be interpreted as band-passes. For example, the difference D_2 between Layer 2 and 3, encodes spectral modulations between 1/2 and 1/4 $\frac{\text{cycles}}{\text{ERB}}$, which are considered important for the recognition of vowels. The difference D_3 between Layer 3 and 4, encodes temporal modulations above $\frac{1}{30}$ ms $\approx$ 33.3 Hz, in bands with a spectral resolution of 4 ERB and are considered to be important for the recognition of consonants. The difference D_5 between Layer 5 and 6, encodes spectral modulations between 1/8 and 1/16 $\frac{\text{cycles}}{\text{ERB}}$ (with a temporal resolution of $\approx$ 33.3 Hz) and are considered not to be crucial for speech recognition.

The differences provide a band-pass decomposition of the logMS in spectral and temporal modulations; hence, Layer 1 can be written as the sum of Layer 7 and D_6+D_5+D_4+D_3+D_2+D_1. However, the differences can also be multiplied by factors F_1, F_2, F_3, F_4, and F_5, which compress or expand the corresponding encoded dynamic. The element-wise summation of the modified differences $F_1 \cdot D_1 + F_2 \cdot D_2 + ... + F_6 \cdot D_6$ gives a modified version of the logMS. The element-wise difference with Layer 1, i.e. the logMS of the unprocessed signal, gives the time- and frequency-dependent gains which are required to perform the desired modification. Only F_2, F_3, F_4, and F_5 were exposed as optimization parameters, while F_1 and F_6 were set to 1.

4) Application of gains of step 1), 2), and 3) to the unprocessed waveform was performed with an overlap-add re-synthesis with 10 ms window shift and 20 ms windows length. For this, the gains of step 1), 2), and 3) were combined, i.e., summed up element-wise, which gives the desired total change in gain. The FFT coefficients of the signal frames, that were already calculated for the logMS, were multiplied with the interpolated (at center frequencies of the respective FFT bin) desired change in gain. After taking the inverse FFT of the modified FFT coefficients, the real parts of the frames were multiplied with Hanning windows of 20 ms duration and merged into a time-signal with overlap-add re-synthesis. The logMS of a processed portion of a clean Spanish speech signal with S=6 dB per octave, C=0 dB per octave[2], L=500 Hz, U=8000 Hz, F_1=1, F_2=3, F_3=2, F_4=1, F_5=0, and F_6=1 is depicted in the lower panel of Figure 1.

2.2. Simulations of speech tests with FADE

With FADE, the approach is to *simulate* a speech test, i.e., a speech recognition experiment with listeners. This is achieved by training and testing a modified ASR system at different SNRs and derive a psychometric function from the achieved recognition performance. The matrix sentence tests employed in the Hurricane Challenge 2.0 are especially well suited for this approach, because they have a simple grammar which can be easily modeled with the employed ASR technology.

The code for the FADE simulations with the Hurricane Challenge 2.0 data is available[2]. Matrix test sentences in three languages (Spanish, English, and German), impulse response for three positions (near, mid, far), and a noise recording were provided by the organizers of the challenge. In the following,

[2]https://doi.org/10.5281/zenodo.3725679

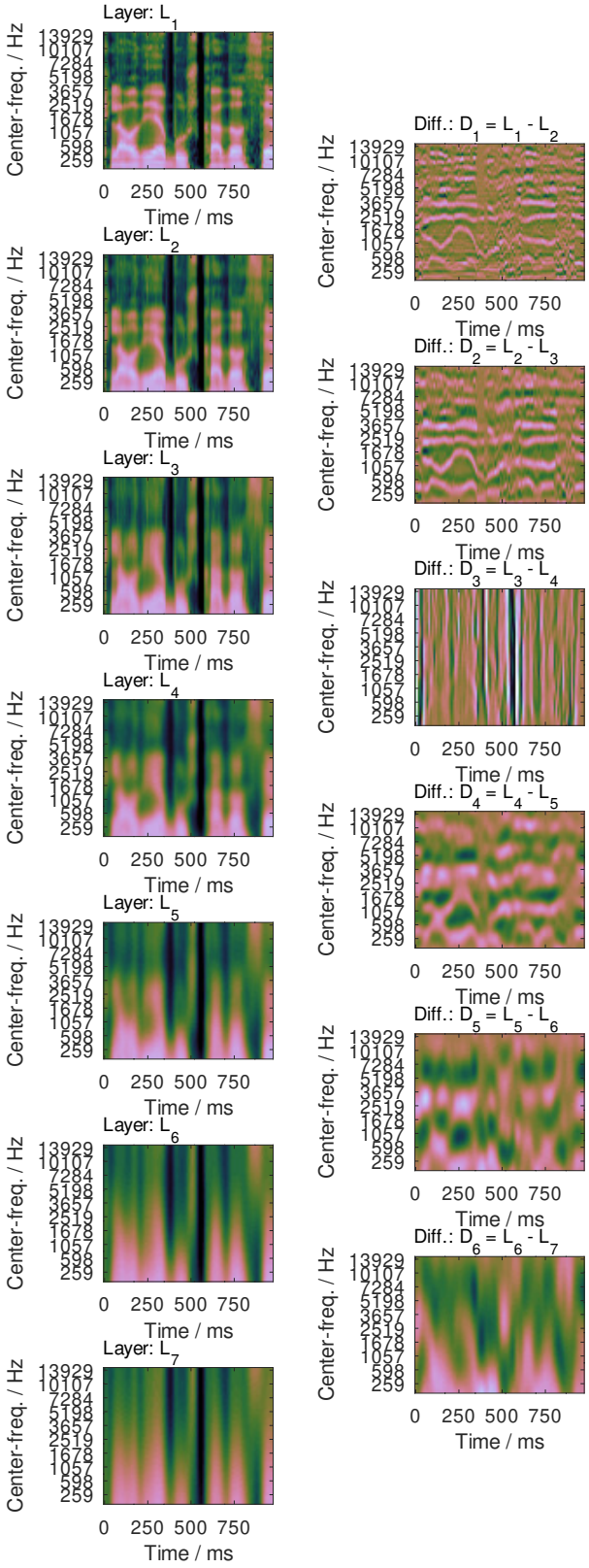

Figure 2: *Spectro-temporally smoothed versions (layers) of the log Mel-spectrograms of the unprocessed clean speech signal (left column) and the element-wise differences between adjacent layers (right column). The color indicates the level, in the left column like in Figure 1, between 0 and 80 dB SPL, in the right column from -10 to 10 dB.*

Table 2: *Initial and optimized parameter values for IISPA.*

Parameter **Language**	S $\frac{\text{dB}}{\text{octave}}$	C $\frac{\text{dB}}{\text{octave}^2}$	F_2	F_3	F_4	F_5
Initial	0	0	1	1	1	1
Spanish	6	0	3	2	1	0
English	3	0	3	2	1	0
German	9	-1	3	2	1	0

the most important steps to simulate a speech test with these signals in FADE are described.

The clean speech signals of the selected language (e.g., Spanish) were processed with IISPA and normalized to the RMS amplitude of the clean signal. Then, the processed signals were convolved with the selected impulse response (e.g., mid) and mixed with random portions of the noise signal at SNRs from -24 dB to 6 dB in 3 dB-steps. A training corpus was generated by repeating this step until a total of 400 sentences per SNR were generated. A test corpus was generated by repeating this step until a total of 100 sentences per SNR were generated. The training corpus was used to train an ASR system for each SNR which can discriminate the 50 words of the matrix sentence test. The test corpus was used to evaluate the recognition performance of the ASR systems. The predicted outcome of the speech test was the lowest SNR at which at least one of the ASR systems achieved the desired recognition performance, e.g., 50% correct for the SRT-50. Pilot experiments showed that the SRT-50 would not be a good choice to optimize a speech-intelligibility improving algorithm, because it allows psychometric functions which never exceed 60% correct recognition rate. Hence, in this work, the SRT-90 was simulated, i.e., the SNR at which 90% of the words could be correctly recognized.

2.3. IISPA parameter optimization with FADE

After pilot simulations, the frequency range was limited by L and U to the range between 500 and 8000 Hz (cf. Figure 1) for all languages. The other parameter were initialized according to Table 2. To optimize the parameters, the following procedure was used. Only the impulse responses for the mid position (as the most representative) were considered during the optimization in order to reduce the number of simulations. For each optimization parameter a set of values was considered. Slopes S from -10 to 0 dB per octave and curvatures C from -4 to 4 dB per octave^2, in 1 dB-steps. For the modulation factors $F_{2..5}$ values of 0, 1, 2, 3, 4, and 5. The SRT-90 for the initial parameter values was simulated. Then all considered parameters were tested in the following order: first all values for the parameter S, then all values for C, all values for F_2 to F_5. Always when a predicted SRT was lower than the currently lowest SRT, the best parameter configuration was updated. If this happened, the optimization was started again after all parameters were iterated. The process finished once no parameter change improved the current best configuration anymore.

2.4. Evaluation of IISPA in the Hurricane Challenge 2.0

IISPA with the corresponding optimized parameter configuration (cf. Table 2) was used to process the provided speech signals, and the processed signals were submitted to the Hurricane Challenge 2.0. There, the signals were convolved with impulse responses which were similar (but not identical) to the ones provided for optimization and mixed at three SNRs (low,

mid, high) with a noise recording similar to (but not identical) to the one provided for optimization. In total, more than 60 listeners per language were presented with the processed and unprocessed stimuli and had to recognize the matrix sentences. The equal-performance SNR improvement (EPSI) according to [9] was calculated from the recognition performance, which was evaluated at three fixed SNRs per condition. It was positive if the processing improved the SNR, and negative otherwise.

3. Results and Discussion

3.1. IISPA parameter optimization with FADE

When the SRT-50 was used for the optimization, it tended to narrow down the bandwidth of the speech signals with the parameters U and L to less than 1 to 3 kHz. There was sufficient information in this frequency range for the ASR system to correctly recognize 50% of the words, but not to achieve recognition rates above 60%, i.e., the simulated psychometric functions were very flat. To avoid this undesired flatting of the psychometric function, the parameters U and L were fixed to 0.5 and 8 Hz, respectively and the SRT-90 was used as the optimization criterion instead of the SRT-50. The SRT-90 might also better reflect realistic listening conditions. But the same optimization for a more complex speech test with than the matrix sentence test might have resulted in different parameters.

Because the simulations have a stochastic component and many good configurations achieved SRT-90s close to each other, the 10 best configurations for each languages were compared for similarities. Finally, the parameters in Table 2 were chosen. The same values for $F_{2..5}$ were chosen for all languages, because they resulted in low simulated SRT-90s. F_2=3 effects an expansion of spectral modulations between between 1/2 and 1/4 $\frac{\text{cycles}}{\text{ERB}}$ by a factor of 3, F_3=2 an expansion of temporal modulations above $\approx$ 33.3 Hz by a factor of 2, and $F_5 = 0$ a removal of spectral modulations between 1/8 and 1/16 $\frac{\text{cycles}}{\text{ERB}}$. These settings effect the spectral and temporal contrast enhancement that can be observed in the lower panel in Figure 1.

Good values for slope S and curvature C turned out to be language (or maybe rather speaker) dependent, where the predicted improvement was approximately 5 dB for all languages. The simulation results underline that, due to the complex, nonlinear nature of the problem, a data-driven optimization for speech recognition performance requires more than an OIM: 1) A meaningful measurement (here the SRT-90 with the matrix sentence test was chosen), 2) a representative set of conditions (here the mid position condition was chosen) 3) and an interpretation of the universality of the best values (here some parameter values were used across languages). Their combination increases the chances to uncover concepts that might generalize to other than the simulated conditions.

3.2. Evaluation of IISPA in the Hurricane Challenge 2.0

The mean recognition performance for languages, positions, and SNRs are presented in Figure 3. The processing achieved an EPSI, i.e., an SNR improvement, of up to 4.8 dB in the Spanish, up to 3.8 dB in the German, and up to 2.1 dB in the English evaluation. The improvements were language-dependent, which deviates from the prediction. Only for the Spanish talker, who coincidentally is the author of this manuscript, the predicted improvement of 5 dB was observed. Also, the improvements depended on the position; the highest were observed for the near or mid positions. Lower improvements were observed for the far position, where the proposed processing slightly decreased the performance with the English talker.

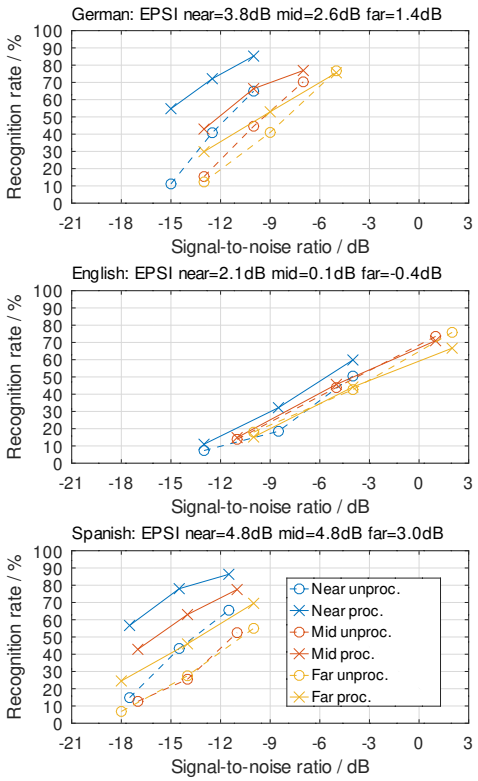

Figure 3: *Average word recognition rates depending on the SNR measured with matrix sentence tests in Spanish, English, and German language, for the near, mid, and far positions.*

That the same optimization strategy results in such variable improvements for the different talkers was unexpected. The FADE simulations did not predict this outcome despite prior work indicating that the approach can predict language effects [5]. Too few data points were available to speculate on the reasons why the processing and optimization with FADE worked better for the Spanish and German talker than for English one. The main difference between the processing for the languages was the spectral slope parameter S, i.e., the processing was very similar. Hence, there was an interaction between the talker/language and the signal processing, which was not explained by the model. However, the predictions were not performed with the binaural impulse responses and noise signals that where used in the evaluation. These simulations could be performed to evaluate the prediction performance of the model.

4. Conclusion

FADE, an objective ASR-based speech recognition model, was successfully used to optimize IISPA, an intelligibility-improving signal processing scheme. The results show that FADE can provided valuable information for the optimization of a speech processing algorithm. However, the large variability in improvements across languages was not predicted and shows the need for thorough listening experiments.

5. Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 352015383 - SFB 1330 A 3.

6. References

[1] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: The Hurricane Challenge," in *INTERSPEECH*, 2013, pp. 3552–3556.

[2] A. ANSI, "S3. 5-1997, Methods for the calculation of the Speech Intelligibility Index," *New York: American National Standards Institute*, vol. 19, pp. 90–119, 1997.

[3] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," *Computer Speech & Language*, vol. 35, pp. 73–92, 2016.

[4] M. R. Schädler, A. Warzybok, S. D. Ewert, and B. Kollmeier, "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2708–2722, 2016.

[5] M. R. Schädler, D. Hülsmeier, A. Warzybok, S. Hochmuth, and B. Kollmeier, "Microscopic multilingual matrix test predictions using an ASR-based speech recognition model," in *INTERSPEECH*, 2016, pp. 610–614.

[6] M. R. Schädler, A. Warzybok, and B. Kollmeier, "Objective prediction of hearing aid benefit across listener groups using machine learning: Speech recognition performance with binaural noise-reduction algorithms," *Trends in Hearing*, vol. 22, p. 2331216518768954, 2018.

[7] M. R. Schädler, A. Warzybok, and B. Kollmeier, "Individual aided speech recognition performance and predictions of benefit for listeners with impaired hearing employing FADE," *Trends in Hearing*, under revision.

[8] J. Rennies-Hochmuth, H. Schepker, M. Cooke, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: The Hurricane Challenge 2.0," in *INTERSPEECH*, 2020, pp. –.

[9] M. R. Schädler and B. Kollmeier, "Separable spectro-temporal gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2047–2059, 2015.