ON SPECTRAL AND TEMPORAL SPARSIFICATION OF SPEECH SIGNALS FOR THE IMPROVEMENT OF SPEECH PERCEPTION IN CI LISTENERS

Benjamin Lentz[1], *Rainer Martin*[1], *Kirsten Oberländer*[2], *Christiane Völter*[2]

[1] Institute of Communication Acoustics Ruhr-Universität Bochum, Bochum, Germany
[2] Department of Otorhinolaryngology, Head and Neck Surgery,
St. Elisabeth-Hospital, Ruhr-Universität Bochum, Bochum, Germany
email: {benjamin.lentz, rainer.martin, kirsten.oberlaender, christiane.voelter}@rub.de

ABSTRACT

The perception of complex signals such as music or speech in noise is a difficult task for most cochlear implant (CI) users. Furthermore, there is a wide variability in speech recognition so that some also face difficulties in everyday situations with only little or no noise. In this study, two methods inspired by music simplification approaches were developed and evaluated through instrumental measures and in listening tests with adult CI listeners. Signals were processed based on a separation into transient and harmonic parts. After sparsification of the harmonic spectrum by principal component analysis (PCA) or by an individualized spectral peak picking approach, the transient and the sparsened harmonic parts were remixed. Significant improvements in speech recognition could be observed when the PCA-based method was applied. This might be caused by noise reduction effects and modulation amplifications as shown by correlation analysis.

Index Terms— Cochlear implants, speech processing, sparsity

1. INTRODUCTION

Although most users strongly benefit from the implantation of cochlear implants (CI) in terms of speech reception in quiet, the perception of complex signals, such as speech in noise or music, remains difficult. This can mostly be attributed to the coarse representation of spectral information at the electrode-nerve interface and channel interactions in the cochlea. Furthermore, as speech reception widely varies among CI users some even face difficulties in quiet environments [1]. Therefore, signal processing algorithms might be an option to improve speech recognition in CI recipients both in quiet and noisy environments.

As most signal processing approaches focus on the removal of noise in a given input signal, their possible benefit on clear speech is limited. Therefore, the goal of our study was to develop a signal pre-processing algorithm for the improvement of speech perception, which should lead to improvements both in noisy or noiseless environments. Our target group are CI listeners who have significant potential for improvement of speech understanding in everyday situations with no or mild environmental noise.

The developed processing approaches are based on music simplification techniques recently proposed to improve the appraisal of classical chamber music [2, 3] or Pop- and Rock music [4] in CI recipients. In the mentioned studies, music signals modified by principal component analysis (PCA) based spectral complexity reduction were preferred over unprocessed signals in CI users [3, 2] or normal hearing listeners in a vocoder CI simulation [4]. Some CI subjects even preferred a strong level of complexity reduction which also sparsened the spectral components of the leading voice [3]. We hypothesize that a spectral sparsification might also improve speech recognition as it has the potential to reduce undesired modulation effects within CI channels and irritations of less important electric stimulation components.

Furthermore, the enhancement of speech onsets might positively influence speech perception in CI users [5]. As the approach of [4] also aims at enhancing the temporal perception of music such as improved audibility of onsets and rhythm, transferring this approach to speech might be beneficial.

The remainder of this paper is organized as follows. In Section 2 the proposed methods and their different components are described in detail. The experimental design to study objective and subjective effects of the methods is described in Section 3. Section 4 describes the experimental results. Conclusions are drawn in Section 5.

2. METHODS

2.1. Signal Processing for Spectral Sparsification

Two signal processing strategies - the Spectral Peak Picking (SPP) algorithm and the Principal-Component-Analysis-based Sparsification (PCAS) - have been developed for the spectral sparsification of speech signals. Both involve the following processing steps. The input signal $x[t]$ with discrete time index t is separated into harmonic and transient signal parts $x_{\mathrm{harm}}[t]$ and $x_{\mathrm{tran}}[t]$. For this, we use an algorithm that has been developed for the harmonic/percussive sound separation of music signals [6]. It provides a threshold parameter β_{tran} which controls the assignment proportions between harmonic and transient components. In our implementation it was chosen such that mostly unvoiced broadband sounds and transitions are contained in the transient signal and harmonic and stationary sounds such as vowels, laterals, nasals, and stationary noise are assigned to the harmonic signal.

After a pre-emphasis is applied on $x_{\mathrm{harm}}[t]$, it is processed using one of two different sparsification approaches described

The study was supported by MED-EL, Innsbruck, Austria.

in the following sections in more detail. These methods emphasize harmonic structures, especially those close to formant regions, while more broadband components or, depending on the degree of sparsification, also less dominant harmonic structures are attenuated. Thus, spectral structures which are assumed to be most important for speech reception are emphasized. Other portions, however, which might rather be detrimental or causing channel interactions are attenuated.

The signal $x_{\text{tran}}[t]$ is not spectrally sparsened because it contains short transitions and noise-like broadband sounds which would not benefit from a selection of frequency components across extended temporal segments. This has been shown in a listening test where the perception of percussive signals was impaired when a PCA-based spectral complexity reduction was applied on a broadband music signal without prior harmonic/percussive separation [4].

After sparsification and de-emphasis, the obtained simplified signal $\tilde{x}_{\text{harm}}[t]$ is mixed with the weighted transient signal part to obtain $\tilde{x}[t] = \tilde{x}_{\text{harm}}[t] + \alpha x_{\text{tran}}[t]$. Here, an amplification factor α can be applied to accentuate fast speech variations such as transients and transitions. Since the transient signal part is not processed, the assignment threshold β_{tran} strongly influences the level of sparsification. We use a setting of $\beta_{\text{tran,agg}} = 1.4$ for the PCAS method which causes larger portions to be assigned to harmonic components and hence be sparsened. For the SPP method we use the setting of $\beta_{\text{tran,mild}} = 1$ whereby less portions are assigned to the harmonic part. We thereby avoid an overly aggressive sparsification through the SPP as we expected it to be more effective than PCAS.

The spectral sparsification can generate a slight reverberation effect because some spectral information is missing to warrant sharp onsets and offsets in the modified signal. To reduce this effect, the sub-bands of the original and processed signals are rectified and low-pass filtered to obtain their respective envelopes $e_{\text{i}}[t]$ and $\tilde{e}_{\text{i}}[t]$. In a non-linear function

$$\tilde{e}_{\text{i,mod}}[t] = \begin{cases} e_i[t] & \text{if } e_i[t] < \tilde{e}_i[t] \\ \tilde{e}_i[t] & \text{otherwise} \end{cases} \tag{1}$$

the respective lower envelope value of $x[t]$ and $\tilde{x}[t]$ is determined and later enforced on $\tilde{x}[t]$. This yields the final output signal $\tilde{x}_{\text{mod}}[t]$ and ensures that pauses, silence, and modulations of $x[t]$ are well maintained.

Examples of the realized signal simplifications are depicted in the spectrograms in Fig. 1. Transients stick out as vertical lines in the speech spectrogram a) and the separated transient part c). The vertical structures are equally retained by the harmonic/ transient sound separation in the PCAS and SPP approach d), and e). Furthermore, formant regions are emphasized and noise is reduced.

2.1.1. Spectral Peak Picking algorithm

In the first processing step of the SPP approach, the pre-emphasized harmonic input signal $x_{\text{harm}}[t]$ is transformed into spectral domain using a short-time Fourier transform (STFT) yielding the spectrogram $X \in \mathbb{R}^{F \times T}$ with a number of F non-redundant frequencies and T time bins. The obtained spectrogram is divided into N_{F} frequency regions corresponding to the sub-bands defined by the individual CI fitting data where usually each sub-band is assigned to one stimulation electrode. In each time bin of the spectrogram with index $\mu \in \{1, \ldots, T\}$, one frequency bin f_{i} per frequency region $i \in \{1, \ldots, N_{\text{F}}\}$ is chosen to be the most important by picking the largest magnitude value of X or its temporarily smoothed version X_{temp}. These frequency bins f_{i} are copied from X to the sparsened spectrogram $\tilde{X}$ while all other frequencies of $\tilde{X}$ are set to zero. Also, the original phase information is added to $\tilde{X}$ so that it can be transformed into the time-domain again. To emphasize the temporal consistency of harmonics, the frequency region

$$\begin{aligned} &X_{\text{temp}}\left[f_{\text{i}} - \frac{N_W}{2} + m, \mu + 1\right] \\ &\qquad = X\left[f_{\text{i}} - \frac{N_W}{2} + m, \mu + 1\right] \cdot (0.5 \cdot w[m] + 1) \end{aligned} \tag{2}$$

of the last selected frequency f_{i} is emphasized in the next time bin $\mu + 1$ by applying a Kaiser window $w \in \mathbb{R}^{N_{\text{W}}}$ with sample index $m \in \{0, \ldots, N_{\text{W}}\}$ across this frequency region. This simple tracking procedure avoids random variations of the selected frequencies from one time bin to the next. Therefore, the choice of frequencies becomes smoother over time which helps to establish continuous harmonics and, in general, helps to avoid cross-modulation effects between harmonic components within specific CI channels.

After sparsification, sub-bands up to a frequency of 3.5 kHz are each normalized to their original power. Furthermore, the broadband sparsened harmonic signal is normalized to the power of the original harmonic signal. We thus ensure that the sparsification does not affect the loudness and spectral balance too much. However, on sub-bands above 3.5 kHz, the selected frequencies are very sparse and a normalization to the original energy would make them overly dominant. Therefore, we do not apply the normalization to these sub-bands.

2.1.2. PCA-based Sparsification

The PCAS approach is inspired by spectral simplification techniques developed for music signals. Its processing steps are explained in detail in [2], [7] and [4]. This approach applies a PCA-based low-rank approximation to blocks of successive spectra of the harmonic signal part to reduce it to the statistically most important spectral components. These are spectral components which are the most dominant over the duration of the block and, because of their high power, particularly well represented in the first N_{Comp} principal components used for approximation. In music, these especially contain harmonic structures from the leading instrument. In speech signals, these might especially contain harmonic structures from vowels and here in turn dominant harmonics close to formant frequencies. Softer harmonics or more broadband sounds, however, are attenuated. The degree of simplification or sparsification can be controlled by the number of principal components. Fewer principal components lead to a larger reduction of spectral components. In this study, we use a number of $N_{\text{Comp}} = 8$ principal components. For processing speech signals the approach is modified by using a STFT instead of a Constant-Q transform [8]. Furthermore, we use a fixed block duration of 100 ms, which is shorter than that used in music signals, because the harmonic structures in speech often vary faster than in music.

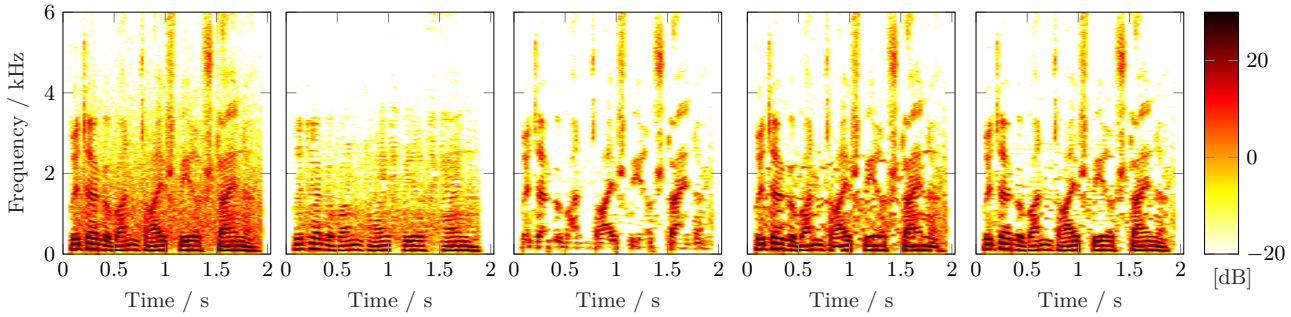


Fig. 1: Spectrograms of a) a speech signal with stationary speech shaped noise of SNR 15 dB, its separated b) harmonic and c) transient signal parts (with separation threshold $\beta_{\mathrm{tran,agg}} = 1.4$) and versions processed through d) SPP and e) PCAS

3. EVALUATION

3.1. Speech Materials

We used sentences from the German speech reception test "Oldenburger Satztest" (OlSa) [9] with a fixed level of stationary speech-shaped noise as input signals. Experiments in [10] have shown that typical SNR levels range from 5 to 8 dB in communication situations outside and between 9 to 14 dB inside the homes of study participants. [11] found an average SNR of 7.4 dB for older adults in noisy situations and [12] found that positive SNRs ranging from 5 to 15 dB cover the majority of communication situations of hearing-impaired listeners. To approximate these findings and to avoid ceiling effects, noise is added to yield an SNR of 8 dB.

We then study these noisy speech signals and the output of SPP and PCAS in terms of instrumental measures and in a listening experiment. The processed signals are normalized to the original power.

3.2. Instrumental Evaluation

To shed light on the effect of the proposed processing steps, different instrumental measures are calculated on the unprocessed and processed signals.

As the sparsification methods are designed to emphasize harmonic components and remove less prominent parts of the spectrum, their effect should be measurable by the degree of harmonicity or sparsity of the spectrum. One instrumental measure which is often used for this purpose is the spectral entropy (see e.g. [13]). The spectral entropy is calculated for each frame of the magnitude spectrogram where the sum across frequency has been normalized to unity. The frame-wise spectral entropy values are subsequently averaged over time. The spectral entropy becomes lower the sparser the signal spectrum is.

Temporal modulation is measured using the ModA measure introduced in [14]. This measure was developed to quantify the amount of reverberation in a signal based on its modulation strength by calculating the area of the modulation spectrum. Higher values of ModA indicate stronger modulation. We use the weighting factors of frequency bands suggested by [14]. In contrast to the initial normalization of signal amplitudes in the range of -0.8 to 0.8 [14], we normalize the signal power to 0.025 which better matches the normalization used in the listening experiments. The value of 0.025 was chosen because it scales the ModA values in a comparable range (0-1) as proposed by [14].

Furthermore, the frequency weighted segmental signal-to-noise ratio (fwsegSNR) is calculated using the implementation provided in [15]. This measure comprises a psychoacoustically motivated frequency weighting of SNR and an averaging of time segments.

The features are calculated also for the intermediate steps of the processing methods, i.e. for the separated harmonic H and transient T parts and after simplification of the harmonic part. In combination with the different separation thresholds $\beta_{\mathrm{tran,agg}} = 1.4$ and $\beta_{\mathrm{tran,mild}} = 1$ this yields the signals H_{Mild}, T_{Mild}, H_{Agg}, and T_{Agg}. Furthermore, the simplified versions of the H part are considered, whereby the versions $H_{\mathrm{simp,SPP}}$ and $H_{\mathrm{simp,PCAS}}$ are generated.

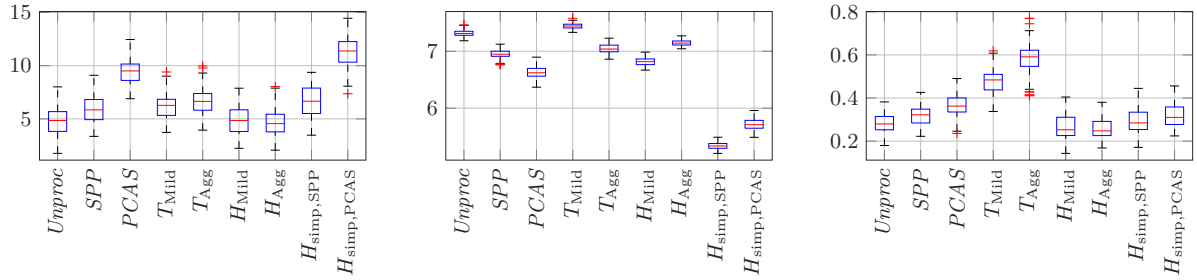
3.3. Listening Experiments

The inclusion criteria for the adult participants (age $\geq$ 18 years) were unilateral or bilateral CI implantation, postlingual hearing loss, CI experience of at least 9 months and no severe cognitive and neurological impairments.

Test signals were presented to only one CI implanted ear at 65 dB SPL acoustically via the Interacoustics TBS25 test chamber. We tested each signal version (Unproc, SPP, PCAS) using the non-adaptive closed OlSa test and three short lists containing 20 sentences each. The first list was used to familiarize the participants with the sound of the signal version under test and was not included in the evaluation. To be sure that the participants understood the procedure and were familiar with the test material, two additional short lists not included in the evaluation were presented at the beginning of the test. The lists were selected pseudo-randomly so that each list was presented at most once to each participant.

4. RESULTS

In Fig. 2 a), the fwsegSNR shows that most noise portions are assigned to the H component leading to a lower fwsegSNR than in the unprocessed (*Unproc*) condition. In T, however, noise is removed so that fwsegSNR improves. Through the comparison of T_{Mild} and T_{Agg}, we can see that this is even more the case for the aggressive threshold setting so that even less noise is contained in T_{Agg} and more in H_{Agg}. The



a) fwsegSNR [dB] b) spectral entropy [nats] c) ModA

Fig. 2: Distribution of signals features for different signal versions and intermediate processing steps.

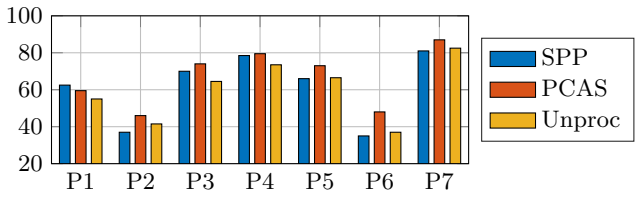


Fig. 3: Speech recognition results (in %) of participants P1-P7 in OlSa tests

large noise portions contained in H also explain the higher spectral entropy values for H_{Agg} than for T_{Agg}. Through the simplification of H, large noise portions are removed, whereby theses effects are reduced. Thus, fwsegSNR increases when it is processed with the simplification methods in $H_{\text{simp,SPP}}$ and $H_{\text{simp,PCAS}}$ and spectral entropy strongly decreases.

The temporal modulation indicated by ModA (see Fig. 2 c)) strongly increases for T compared to *Unproc*. This shows the importance of the structures contained in T. A clear difference between T_{Mild} and T_{Agg} can be seen. This is certainly the case because more noise portions are assigned to T_{Mild} which disturb the accentuation. The H component show lower ModA values than *Unproc*. As H rather contains stationary sounds and, as the fwsegSNR indicates, also the majority of noise, modulations are certainly attenuated. The modulations increase again after H is sparsened in $H_{\text{simp,SPP}}$ and $H_{\text{simp,PCAS}}$.

Considering only the signals presented in the listening tests (*Unproc*, *SPP*, *PCAS*) an improvement compared to the unprocessed speech can be seen for both processing methods in all measures. The best improvements are obtained for *PCAS*.

Results of the exploratory listening test are depicted in Fig. 3. The individual speech reception outcomes range from 35% to 87% and are highly variable among the CI listeners. Improvements in speech recognition can be observed for all participants when using the PCAS processing. An ANOVA applied on the recognition differences between PCAS processed and unprocessed speech indicates a significant ($p<0.005$) improvement with a mean benefit of 6.5%. For SPP, speech recognition results are more mixed so that an improvement can be observed for three participants and a slight deterioration for four participants. Overall the differences (average benefit of 1.36%) to unprocessed speech are not significant ($p>0.1$).

To investigate correlations between signal features and test outcomes among all participants, we calculate delta features $\Delta z_{\text{k,s,p}}$ for each sentence presented to a participant. Here s is the index of sentences, k of processing methods and p of participants. The delta features are calculated as difference $\Delta z_{\text{k,s,p}} = z_{\text{k,s,p}} - \bar{z}_{\text{unproc,p}}$ between features of the respective processing method and mean feature $\bar{z}_{\text{unproc,p}} = \frac{1}{S}\sum_{s=0}^{S} z_{\text{unproc,p,s}}$ averaged over the unprocessed sentences. The same calculation was applied on the speech recognition values to derive delta recognition outcomes for each sentence. Then, correlations between the delta features and delta outcomes were determined across all participants. Using the delta values, offsets between the outcome levels of different participants are compensated and common correlations can be studied more reliably.

The most prominent yet weak correlation ($r=0.19$, $p<0.001$) of test outcomes and instrumental measures occurs for fwsegSNR. It shows, that speech recognition was improved when noise portions were removed. This is in line with studies showing a positive effect of noise reduction on speech recognition of CI listeners [16]. The sparsification of speech signals was not a significant factor ($r=0.01$, $p>0.1$). Furthermore, a significant correlation occurs between delta speech recognition scores and delta ModA ($r=0.16$, $p<0.001$). This correlation indicates, that a stronger modulation could be beneficial for speech intelligibility.

5. DISCUSSION AND CONCLUSION

While the SPP approach did not have a significant impact, the PCAS method did significantly improve speech recognition in adult CI users. According to correlation analysis, this might be due to the effect of noise reduction and the amplification of speech modulation. For the spectral sparsification, however, no significant effect was observed. While for music signals the reduction of weaker harmonic overtones helps to improve the music appreciation in CI listeners this does not seem to be the case for speech signals. This might be because music perception relies more on pitch and therefore clear harmonic structures. Speech reception, however, is based on coarser spectral structures such as formants.

As the principles of the algorithm are not limited to the reduction of noise but based on emphasizing the most important speech structures, they are promising for applications on clear speech. Here, they might help to simplify speech signals for poorly performing CI users. Especially the emphasis of speech modulations seems suitable for this purpose. As the modulations are mostly retained in the transient signal part, the harmonic / transient sound separation plays an important role and might be a suitable and blind alternative to envelope enhancement approaches such as [5].

6. REFERENCES

[1] Laura K. Holden, Charles C. Finley, Jill B. Firszt, Timothy A. Holden, Christine Brenner, Lisa G. Potts, Brenda D. Gotter, Sallie S. Vanderhoof, Karen Mispagel, Gitry Heydebrand, and Margaret W. Skinner, "Factors affecting open-set word recognition in adults with cochlear implants," *Ear Hear.*, vol. 34, no. 3, pp. 342–360, 2013.

[2] Anil Nagathil, Claus Weihs, Katrin Neumann, and Rainer Martin, "Spectral complexity reduction of music signals based on frequency-domain reduced-rank approximations: An evaluation with cochlear implant listeners," *J. Acoust. Soc. Am. (JASA)*, vol. 142, no. 3, pp. 1219–1228, 2017.

[3] Johannes Gauer, Anil Nagathil, Rainer Martin, Jan Peter Thomas, and Christiane Völter, "Interactive Evaluation of a Music Preprocessing Scheme for Cochlear Implants Based on Spectral Complexity Reduction," *Front. Neurosci.*, vol. 13, pp. 1206, nov 2019.

[4] Benjamin Lentz, Anil Nagathil, Johannes Gauer, and Rainer Martin, "Harmonic / percussive sound separation and spectral complexity reduction of music signals for cochlear implant listeners," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2020, pp. 8713–8717.

[5] Raphael Koning and Jan Wouters, "Speech onset enhancement improves intelligibility in adverse listening conditions for cochlear implant users," *Hearing Research*, vol. 342, pp. 13–22, 2016.

[6] Jonathan Driedger, Meinard Müller, and Sascha Disch, "Extending Harmonic-Percussive Separation of Audio Signals," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2014, pp. 611–616.

[7] Anil Nagathil, Claus Weihs, and Rainer Martin, "Spectral complexity reduction of music signals for mitigating effects of cochlear hearing loss," *IEEE/ACM Trans. Audio Speech and Lang. Process.*, vol. 24, no. 3, pp. 445–458, 2016.

[8] Judith C Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Am. (JASA)*, vol. 89, no. 1, pp. 425–434, 1991.

[9] Birger Kollmeier and Matthias Wesselkamp, "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," *J. Acoust. Soc. Am. (JASA)*, vol. 102, no. 4, pp. 2412–2421, 1997.

[10] Karl S Pearsons, Ricarda L Bennett, and Sanford A Fidell, *Speech levels in various noise environments*, Office of Health and Ecological Effects, Office of Research and Development, US EPA, 1977.

[11] Yu Hsiang Wu, Elizabeth Stangl, Octav Chipara, Syed Shabih Hasan, Anne Welhaven, and Jacob Oleson, "Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss," *Ear Hear.*, vol. 39, no. 2, pp. 293–304, 2018.

[12] Karolina Smeds, Florian Wolters, and Martin Rung, "Estimation of signal-to-noise ratios in realistic sound scenarios," *J. Am. Acad. Audiol.*, vol. 26, no. 2, pp. 183–196, 2015.

[13] Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky, "Spectral entropy based feature for robust ASR," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2004, vol. 1, pp. 193–196.

[14] Fei Chen, Oldooz Hazrati, and Philipos C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 311–314, 2013.

[15] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2007.

[16] Fergal Henry, Martin Glavin, and Edward Jones, "Noise reduction in cochlear implant signal processing: A review and recent developments," *IEEE reviews in biomedical engineering*, 2021.