SPEECH ARTIFACT REMOVAL FROM EEG RECORDINGS OF SPOKEN WORD PRODUCTION WITH TENSOR DECOMPOSITION

Holy Lovenia[1,3], Hiroki Tanaka[1], Sakriani Sakti[1,2], Ayu Purwarianti[3], and Satoshi Nakamura[1,2]

[1] Nara Institute of Science and Technology, Japan
[2] RIKEN, Center for Advanced Intelligence Project AIP, Japan
[3] Department of Informatics, Bandung Institute of Technology, Indonesia

holy.lovenia@gmail.com, {hiroki-tan, ssakti, s-nakamura}@is.naist.jp, ayu@stei.itb.ac.id

ABSTRACT

Research about brain activities involving spoken word production is considerably underdeveloped because of the undiscovered characteristics of speech artifacts, which contaminate electroencephalogram (EEG) signals and prevent the inspection of the underlying cognitive processes. To fuel further EEG research with speech production, a method using three-mode tensor decomposition (time x space x frequency) is proposed to perform speech artifact removal. Tensor decomposition enables simultaneous inspection of multiple modes, which suits the multi-way nature of EEG data. In a picture-naming task, we collected raw data with speech artifacts by placing two electrodes near the mouth to record lip EMG. Based on our evaluation, which calculated the correlation values between grand-averaged speech artifacts and the lip EMG, tensor decomposition outperformed the former methods that were based on independent component analysis (ICA) and blind source separation (BSS), both in detecting speech artifact (0.985) and producing clean data (0.101). Our proposed method correctly preserved the components unrelated to speech, which was validated by computing the correlation value between the grand-averaged raw data without EOG and cleaned data before the speech onset (0.92-0.94).

Index Terms— Speech artifact removal, tensor decomposition, EEG, spoken word production

1. INTRODUCTION

Many cognitive processes are involved in the act of speaking, such as phonetic encoding and articulation [1], even when the utterances are simply naming a common object, e.g., dog [2]. To uncover the representations and brain processes underlying word production in speech, the inspection of cognitive processes in word production can be supported by tools, including an EEG. An EEG can obtain a high temporal resolution [3], which enables research of brain activity with a precise time course. Unfortunately, research in EEG-related fields involving such overt speech as language and word production suffers from noise caused by non-brain activities.

EEG experiments always encounter problems because of its low signal-to-noise ratio (SNR). Even without speaking, only eye blink, the SNR is already low.

Moreover, many muscle movements are required to produce speech, e.g., mouth, jaw, and tongue movements [3]. As a result, the cognitive components of spoken word production processes are inevitably contaminated by these muscle movements [2], which will be addressed as speech artifacts below. Overlapped by speech artifacts, it is difficult to tell whether the components from the recorded data are real brain activities or merely noise. To avoid the occurrence of speech artifacts, many works on EEG data have experimented with covert speech planning and delayed speaking tasks. Over the years, covert speech experiments have undeniably sparked essential discoveries by EEG studies. However, tasks using immediate overt speech are preferred in several situations. In overt speech production, no altering or omitting of the cognitive steps (e.g. articulation) involved in speech production is needed because speech is produced naturally. Hence, speech artifact removal methods must free EEG data from speech-related artifacts during overt speech production tasks. Vos et al. utilized BSS and canonical correlation analysis to separate the muscle movements regarding speech from the raw data [4]. A more recent study by Porcaro et al. exploited ICA to attenuate speech artifacts and investigated the original sources that are related to the act of speaking [2]. Both methods are based on matrix decomposition [2][4].

EEG recordings normally have at least two modes, time and space. The channel (or component) is represented by the space mode, and the brain activity across time is described by the time mode. In matrix decomposition methods, such as BSS and ICA used in previous works, both of these modes extract the origin sources of brain activity. However, besides time and space mode, there can be more modes involved in EEG experiments, such as frequency, condition, group, and subject. Most signal-processing methods only support two-way data instead of three or more dimensions. Thus, efforts have tried to fit all of the required modes into the two-dimensional EEG data. Various attempts have unfolded a

multi-way array into a matrix, including concatenation and stacking to the space/time mode [5]. A tensor is a multi-way array that is regarded as high-order when it has more than two modes. EEG observation of the temporal, spectral, spatial, and other changes in brain activities can be done simultaneously with a tensor [5]. Tensor factorization also enables the consideration of more than two modes during decomposition [6]. The present study has two main objectives. The first is to propose a new method for removing speech artifacts. The proposed method is based on tensor decomposition for sources reconstruction. The second is to evaluate the performance of the proposed method (using tensor decomposition) with the existing methods (using ICA and BSS). Then we validated the cleaned data obtained by all of the methods to ensure the preservation of brain signals.

2. PREVIOUS METHODS TO REMOVE SPEECH ARTIFACTS

The baselines used in this research came from an adapted method of SAR-ICA by Porcaro et al. [2] and the automatic version of BSS-CCA by Vos et al. [4].

2.1. Speech artifact removal by independent component analysis (SAR-ICA)

SAR-ICA, which is a speech artifact removal (SAR) method that uses ICA for decomposing brain sources [2], is comprised of four main phases: decomposition, artifact detection, control cycle, and clustering. The sources reconstruction used fastICA, and the corresponding independent components (ICs) were marked as artifacts based on their statistical and spectral characteristics, with the electro-oculogram (EOG) and electromyogram (EMG) channels as the benchmarks. In addition, we also visually inspected the averaged trials, the single trials, and the topographical distributions of the ICs to manually categorize whether the corresponding IC was an artifact. After the identification of both the eye-related (ocular) and speech-related artifacts, we applied a control cycle. This step confirmed that the differences between the raw and the artifact-free data were only caused by the artifact removal. If any brain activity was present in the removed data, we reduced the thresholds used for the artifact identification and repeated the corresponding step. Then all of the components were clustered based on their result from the previous steps. After the cluster membership, each cluster was back-projected to the channel space.

2.2. Blind source separation with canonical correlation analysis (BSS-CCA)

As input data used for the BSS-CCA method, we used Cluster 3 (raw data without EOG) by SAR-ICA, which consisted of raw data without ocular artifacts. As its name implies, this method based its decomposition process on BSS [4]. Instead of assuming statistical independence, BSS-CCA uses a different basis to separate the brain sources: autocorrelation properties by canonical correlation analysis (CCA). This method assumes that EMG activity is weakly autocorrelated over time, and brain activity is more likely to be autocorrelated because of its coherence over time. There is another criterion to select the components related to muscle movements. If the average power of the components in the EMG band (approximated by 15-30 Hz) equals or exceeds 1/n (default n is 7) of the average power in the EEG band (approximated by 0-15 Hz), the corresponding component is marked as a muscle (speech-related) artifact. By defining a quantitative criterion to mark the speech artifacts, no visual inspection was needed in BSS-CCA.

3. PROPOSED METHOD USING TENSOR DECOMPOSITION

Unlike the previous methods that used matrix decomposition, the proposed method carried out the sources reconstruction step using tensor decomposition, which is also generated from blind source separation [5]. However, the concept behind these two methods is essentially different. A brief overview of the differences between the baselines and the proposed method is represented in Table 1. In the proposed method, we applied third-order canonical polyadic (CP) decomposition from the Tensor Toolbox to the original data [7][8].

Table 1. Overview comparison of SAR-ICA [2], BSS-CCA [4], and proposed method with tensor decomposition (TD)

	SAR-ICA	BSS-CCA	TD
Decomposition method	ICA	BSS	CPD
EMG channels	Yes	No	Yes
Needs visual inspection	Yes	No	Yes

In CP decomposition (CPD), which is also called canonical decomposition (CANDECOMP) or parallel factor analysis (PARAFAC) [5], the number of components for each mode is the same [5][9] (Fig. 1). When the original data are decomposed by CPD with time, frequency, and channel modes, then the number of temporal, spectral, and space components is identical. Given Nth-order tensor X, CPD is generally described as

$$X = \sum_{r=1}^{R} u_r^{(1)} \bullet u_r^{(2)} \bullet \cdots \bullet u_r^{(N)} + E = \sum_{r=1}^{R} X_r + E$$
$$= \hat{X} + E \approx \hat{X}, \tag{1}$$

where r is the number of components, $n = 1, 2, \cdots, N$ is the mode, and E is the residual [5].
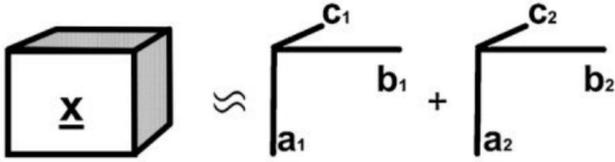
Fig. 1. Third-order tensor with two CPD components

The number of extracted components for each subject was determined by the difference of the fit (DIFFIT) method by Timmerman et al. with a lower bound of 5 and an upper bound of 26 [5][10]. We chose five to prevent the number of components from being too low and 26 EEG channels. The number of extracted components used by the matrix decomposition was also beneficial when applying CP decomposition, such as ICA, which uses the number of sources in the brain [5]. The average number of extracted components from the EEG data across the subjects was approximately 15. The lowest estimation was 8, and the highest number of extracted components was 24. The artifact detection was based on three aspects: the frequency domain characteristics, the time domain characteristics, and a visual inspection. The frequency domain characteristics were observed by calculating the correlation coefficients between the spectral components of the EEG data and the spectral component of the EMG data.

Similar observations analyzed the time domain characteristics. However, there was a significant differentiator: the usage of speech onset. The Pearson correlation values between the temporal components of the EEG data and the temporal component of the EMG data were calculated starting from the averaged speech onset of the corresponding subject until the next second or the end of the trial, whichever happened first. If the correlation coefficient showed a strong association between both the temporal and spectral components and the EMG data, the corresponding Kruskal tensor (the CPD decomposition result of the original tensor [7]) was marked as a speech artifact. We manually inspected the topographical distributions in the space mode and averaged the trial components. Then we based the decision whether to mark the Kruskal tensor as a speech artifact on all of the results.

The difference between before and after removing the Kruskal tensors was marked as speech artifact and observed as a post-removal validation. If a brain signal was caught in the removed Kruskal tensor, then the artifact detection step was repeated. After the previous steps were done, the Kruskal tensors were classified into three different clusters. Cluster 1 (raw data without ocular artifact) contained the average of all the Kruskal tensors. Cluster 2 (speech artifact) was composed of the Kruskal tensors marked as muscle-contaminated. Cluster 3 (cleaned data) consisted of artifact-free Kruskal tensors.

4. EXPERIMENT

4.1. Participants and materials

Nine native Japanese speakers (mean age 23.3, SD 2.6, 8 males, 2 females) with normal or corrected-to-normal vision participated in the experiment. Eight speakers were right-handed, and the rest were corrected-to-right. One subject was removed from further analysis due to a high rate of errors. We used 45 different line drawings of common objects as the stimuli [11].

4.2. Picture-naming task

The raw data collection consisted of pre-experiment and experiment phases. In the pre-experiment phase, the subjects looked at a booklet of picture-name pairs and task description. This phase confirmed that the participants knew the Japanese names of the picture stimuli used in the experiment. The subjects were instructed to only blink after naming the displayed stimulus and to avoid body and head movements as much as possible. There were two blocks in the experiment phase: practice and experiment.

Each block was composed of 45 trials. A rest period was provided between the blocks. In this period, the subjects rested their eyes and body before moving to the next part of the experiment. Each trial randomly used one of the line drawings as the picture stimulus. Throughout the block, each picture appeared once in a random order. At the beginning of every trial, a fixation cross appeared for one second. Then it was replaced with a picture stimulus. The participant named the displayed stimulus as quickly and as accurately as possible. The stimulus remained on display for three seconds. The stimulus presentation was coded using Presentation (Neurobehavioral Systems).

4.3. EEG, EMG, and EOG recordings

The EEG data were recorded with ActiCAP as the EEG cap and BrainAmp DC (BrainProducts) as the amplifier. During our experiment, 27 electrodes were used as EEG channels on the scalp and one as the default reference. Two electrodes monitored the electro-oculogram (EOG), also known as eye artifact (such as eye blinks), just above and below the left eye. A lip electromyogram, which is a speech artifact (muscle movements caused by speech), was recorded by placing two electrodes at the left orbicularis oris superior (OOS) and the left orbicularis oris inferior (OOI) halfway between the center and the corner of the mouth.

5. DATA ANALYSIS

The data analysis was done using the FieldTrip [12] and Tensor Toolbox [7].

5.1. Data pre-processing

The continuous raw data collected from the picture-naming experiment were epoched into trials that began at -1 second and ended at 3 seconds. We removed from further analysis trials that contained incorrect names, missed answers, or self-repair attempts and performed baseline correction and 0.1-30 Hz bandpass filtering. EEG channels were offline re-referenced to the average of the left and right mastoids, and the EOG and EMG data used bipolar derivation as a reference. All of the data were offline resampled to 512 Hz.

5.2. Evaluation

We compared the performances of the methods by calculating the Pearson correlation (R) between the grand-average clusters (speech artifact and the cleaned data) and the lip EMG for 0-1350 ms in the time domain. We assumed that the lip EMG represented the real speech artifact. The clusters were averaged across the subjects. The aim of further validation is to confirm whether the cleaned data cluster preserved the brain signals. Validation computed the correlation values between Cluster 1 (raw data without ocular artifacts) and Cluster 3 (cleaned data) by the proposed method for 0-700 ms (before the earliest speech onset) and 0-900 ms (before the grand-average of the speech onsets). If no brain signal was removed in the cleaned data, they are supposedly highly correlated.

6. RESULTS AND DISCUSSION

The table below presents the detailed results of the evaluation step, which is the absolute Pearson correlation value between the speech artifacts and the cleaned data with the lip EMG in the time domain for 0-1350 ms. A higher correlation value is better for the speech artifact, because it shows an association with the lip EMG. For the cleaned data, a low correlation value is better. Proposed method is written as CPD in bold.

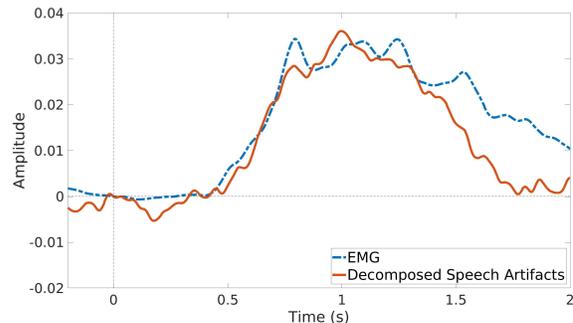**Table 2**. Grand-average correlation to lip EMG (all, $p < 0.01$)

Grand-average data	R (0-1350 ms)
SAR-ICA's Cluster 2 (Speech Artifacts)	0.875
CPD's Cluster 2 (Speech Artifacts)	**0.985**
SAR-ICA's Cleaned Data	0.351
BSS-CCA's Cleaned Data	0.413
CPD's Cleaned Data	**0.101**

As shown in Table 2, the proposed method using tensor decomposition outperformed its baselines, SAR-ICA and BSS-CCA, both in detecting speech artifacts (0.985) and producing cleaned data (0.101). Fig. 2 shows that at most of the time points, the speech artifacts were almost identical to

the lip EMG. However, only a slight difference remains between them, which might be caused by the inability of EMG channels to pick up every single movement related to speech.

Table 3. Grand-average correlation to raw data without EOG before speech onset ($p < 0.01$)

Grand-average data	R (0-700 ms)	R (0-900 ms)
CPD's Cleaned Data	0.927	0.942

Fig. 2. Comparison of normalized grand-average lip EMG and decomposed speech artifacts by proposed method

This difference was also expressed in the validation result in Table 3, which yielded a correlation value of approximately 0.92-0.94 between the raw data without ocular artifact and the cleaned data (CPD) before the earliest speech onset (700 ms) and the average of the speech onsets (900 ms), which indicates that the quality of the cleaned data is enough for subsequent EEG processing. The validation result shows a small difference during the pre-speech onset, which might also be caused by removing the ocular artifact.

7. CONCLUSION

We proposed a speech artifact removal method using tensor decomposition. Our proposed method surpassed the former methods (SAR-ICA and BSS-CCA) both in identifying speech artifacts (0.985) and producing cleaned data (0.101), validated by around 0.92-0.94. Both the evaluation and validation were calculated with the Pearson correlation coefficient using the grand-average of the related clusters and the lip EMG. Since the present study relied on visual inspection, future research should fully automate all of the steps needed for artifact identification.

8. ACKNOWLEDGEMENT

9. REFERENCES

[1] A. Valente, A. Brki, and M. Laganaro, "ERP correlates of word production predictors in picture naming: a trial by trial multiple regression analysis from stimulus onset to response," *Frontiers in neuroscience*, vol. 8, no. 390, 2014. [Online]. Available: http://europepmc.org/articles/PMC4255522

[2] C. Porcaro, M. T. Medaglia, and A. Krott, "Removing speech artifacts from electroencephalographic recordings during overt picture naming," *NeuroImage*, vol. 105, pp. 171 – 180, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811914008751

[3] S. Luck, *An Introduction to the Event-Related Potential Technique*, ser. A Bradford Book. MIT Press, 2014. [Online]. Available: https://books.google.co.id/books?id=SzavAwAAQBAJ

[4] M. de Vos, D. Maarten Vos, S. Ries, K. Vanderperren, B. Vanrumste, F.-X. Alario, S. Huffel, S. Huffel, and B. Burle, "Removal of muscle artifacts from eeg recordings of spoken language production," *Neuroinformatics*, vol. 8, pp. 135–150, 06 2010.

[5] F. Cong, Q.-H. Lin, L.-D. Kuang, X.-F. Gong, P. Astikainen, and T. Ristaniemi, "Tensor decomposition of eeg signals: A brief review," *Journal of Neuroscience Methods*, vol. 248, pp. 59 – 69, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165027015001016

[6] M. Hayato, H. Tanaka, S. Sakti, and S. Nakamura, "Graph regularized tensor factorization for single-trial eeg analysis," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 846–850, 4 2018.

[7] B. W. Bader, T. G. Kolda *et al.*, "Matlab tensor toolbox version 3.0-dev," Available online, Oct. 2017. [Online]. Available: https://www.tensortoolbox.org

[8] B. W. Bader and T. G. Kolda, "Algorithm 862: MATLAB tensor classes for fast algorithm prototyping," *ACM Transactions on Mathematical Software*, vol. 32, no. 4, pp. 635–653, Dec. 2006.

[9] A. Cichocki, D. P. Mandic, A. H. Phan, C. F. Caiafa, G. Zhou, Q. Zhao, and L. D. Lathauwer, "Tensor decompositions for signal processing applications from two-way to multiway component analysis," *CoRR*, vol. abs/1403.4462, 2014. [Online]. Available: http://arxiv.org/abs/1403.4462

[10] M. E. Timmerman and H. A. L. Kiers, "Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima," *British Journal of Mathematical and Statistical Psychology*, vol. 53, no. 1, pp. 1–16, 2000. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1348/000711000159132

[11] T. Nishimoto, K. Miyawaki, T. Ueda, Y. Une, and M. Takahashi, "Japanese normative set of 359 pictures." *Behavior research methods*, vol. 37 3, pp. 398–416, 2005.

[12] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "Fieldtrip: Open source software for advanced analysis of meg, eeg, and invasive electrophysiological data," in *Comp. Int. and Neurosc.*, 2011.

[13] G. Barbati, C. Porcaro, F. Zappasodi, P. M. Rossini, and F. Tecchio, "Optimization of an independent component analysis approach for artifact identification and removal in magnetoencephalographic signals," *Clinical Neurophysiology*, vol. 115, no. 5, pp. 1220 – 1232, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1388245703004784

[14] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, September 2009.

[15] A. G. Asuero, A. Sayago, and A. G. Gonzlez, "The correlation coefficient: An overview," *Critical Reviews in Analytical Chemistry*, vol. 36, no. 1, pp. 41–59, 2006. [Online]. Available: https://doi.org/10.1080/10408340500526766

[16] L. Ganushchak, I. Christoffels, and N. Schiller, "The use of electroencephalography in language production research: A review," *Frontiers in Psychology*, vol. 2, p. 208, 2011. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2011.00208

[17] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Inf. Process. Lett. Rev*, vol. 6, 11 2004.