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

Speech Spectrogram Estimation from Intracranial Brain Activity using a Quantization Approach

Miguel Angrick[1], *Christian Herff*[2], *Garett Johnson*[3], *Jerry Shih*[4], *Dean Krusienski*[5], *Tanja Schultz*[1]

[1]Cognitive Systems Lab, University of Bremen, Bremen, Germany
[2]School for Mental Health and Neuroscience, Maastricht University, The Netherlands
[3] Biomedical Engineering, Old Dominion University, Norfolk, VA, USA
[4] Epilepsy Center, UC San Diego Health, San Diego, CA, USA
[5]ASPEN Lab, Biomedical Engineering, Virginia Commonwealth University, Richmond, VA, USA
miguel.angrick@uni-bremen.de

Abstract

Direct synthesis from intracranial brain activity into acoustic speech might provide an intuitive and natural communication means for speech-impaired users. In previous studies we have used logarithmic Mel-scaled speech spectrograms (logMels) as an intermediate representation in the decoding from ElectroCorticoGraphic (ECoG) recordings to an audible waveform. Mel-scaled speech spectrograms have a long tradition in acoustic speech processing and speech synthesis applications. In the past, we relied on regression approaches to find a mapping from brain activity to logMel spectral coefficients, due to the continuous feature space. However, regression tasks are unbounded and thus neuronal fluctuations in brain activity may result in abnormally high amplitudes in a synthesized acoustic speech signal. To mitigate these issues, we propose two methods for quantization of power values to discretize the feature space of logarithmic Mel-scaled spectral coefficients by using the median and the logistic formula, respectively, to reduce the complexity and restricting the number of intervals. We evaluate the practicability in a proof-of-concept with one participant through a simple classification based on linear discriminant analysis and compare the resulting waveform with the original speech. Reconstructed spectrograms achieve Pearson correlation coefficients with a mean of r=0.5 ± 0.11 in a 5-fold cross validation.

Index Terms: neural signals for spoken communication, speech synthesis, electrocorticography, BCI

1. Introduction

Speech is the first and foremost means of human communication. Millions of people worldwide suffer from severe speech disorders, in particular, those who have completely lost their ability to speak due to neurological diseases like amyotrophic lateral sclerosis (ALS), brain stem stroke, or severe paralysis. For example, ALS can lead to the locked-in syndrome - a state in which the affected are fully conscious and aware of their environment, but have no possibility to produce speech. The most promising technology for restoring the ability to communicate for these individuals are biosignal-based spoken communication systems [1], and more specifically Brain-Computer Interfaces (BCIs) [2].

Various systems targeting speech-impaired users have been developed over the last 30 years [3]. While spelling devices based on the P300 signal have demonstrated some degree of success using non-invasive measurements of brain activity (e.g. electroencephalography) [4, 5], the spelling rate is insufficient for a natural spoken communication. However, using intracranial brain signals to directly control a speech synthesis system in real-time has great potential for a neuroprosthetic spoken communication device.

Several studies have tackled the problem of a direct conversion towards an acoustic speech signal based on brain activity acquired during speech production tasks. In a pilot study, Herff et al. [6] synthesized an acoustic speech waveform by introducing a two-step approach: (1) Prediction of a spectrogram for the spoken utterance using linear regression models and (2) Phase information recovery through the application of the Griffin-Lim algorithm [7]. Results indicate that original and reconstructed waveform reveal significant correlations. To improve the quality and intelligibility of the synthesized speech output, systems based on deep learning methods have recently come into focus. Anumanchipalli et al. [8] applied recurrent neural networks to estimate kinematic trajectories of articulatory movements, which can be decoded into acoustic speech. In another study, we used convolutional neural networks to first estimate the speech spectrogram and employ a WaveNet model as a second step for waveform generation conditioned on these spectral features [9]. Furthermore, Akbari et al. [10] proposed deep neural networks in a speech perception task, for a non-linear regression onto an acoustic representation for subsequent resynthesis. In addition, Herff et al. employed a unit selection approach to reconstruct the acoustic signal by selecting the closest speech unit based on the cosine similarity of the corresponding neural features [11].

The results of the aforementioned studies indicated that regression approaches are suitable methods for the mapping of brain activity data onto an internal representation for resynthesis. However, regression tasks are unbounded and thus may cause unintended amplitude spikes and unnatural increases of the output volume of synthesized speech signals in case the input signals show large variation. Unfortunately, such large fluctuations are common in neural recordings. While amplitude spikes are not a major issue in offline analysis, they could be prohibitively distracting to the user for closed-loop speech decoding.

In this study, we examine the mapping from brain activity features to intermediate representations as a classification task to avoid the unbounded behaviour of regressions. Here, we focus our investigation on two distinct quantization approaches to discretize the continuous space of logarithmic Mel-scaled spectral features into a manageable number of disjoint intervals. To evaluate the practicability of discretized intervals of logMel features for the speech decoding tasks, we employ a simple classification based on linear discriminant analysis (LDA) to select

http://dx.doi.org/10.21437/Interspeech.2020-2946

the appropriate interval. The final acoustic waveform is then reconstructed by applying an iterative approximation on the selected LogMel intervals.

For a proof-of-concept of the proposed method, we rely on time-aligned data from a single participant. The data consists of parallel recordings of ECoG and acoustic signals of prompted speech. For evaluation, we compare both methods with the best-case quantization error for a limited amount of intervals.

2. Material and Methods

2.1. Experiment Setup and Data Acquisition

For our analysis, we relied on a pre-recorded corpus of time-aligned ECoG and speech data acquired during an audible speech production task and focused in this feasibility study on the most promising participant based on overall signal quality and our previous investigation [12]. In each trial, English sentences were first prompted by simultaneous display on a computer screen and a narration via loudspeakers. The participant was then asked to recite the sentence from memory immediately after its visual and auditory presentation. All sentences (50 in total) were taken from the Harvard sentences corpus [13], which provides a phonetically-balanced set of phones.

The participant (age 24, male) in this study was being treated for intractable epilepsy and underwent a medical evaluation to localize the seizure foci prior to surgical resection. For this monitoring process, an electrode grid with 56 channels, a strip with 4 channels and 4 depth electrodes were implanted on the left hemisphere solely based on his clinical needs, while covering some relevant areas for speech production. The participant gave written informed consent and the experiment was approved by an IRB of Mayo Clinic, University of California San Diego and Old Dominion University.

We acquired ECoG signals using stacked g.USB amplifiers (Guger Technologies, Austria) at a sampling rate of 1200 Hz. Acoustic recordings of the participant's speech were done with a Snowball iCE microphone (Blue Microphones, California) using a sampling rate of 48 kHz. All data recordings were synchronized using the general-purpose BCI2000 system [14].

2.2. Feature Extraction

As meaningful features, we focused on the broadband gamma (70-170 Hz) band for the ECoG signals, which is known to contain correlates relevant to speech [15, 16] and language [17] processes. Following best practice in our previous work [6] we removed the linear trend from each ECoG channel and downsampled the ECoG signals to 600 Hz. All channels were re-referenced to a common average (CAR spatial filtering). In order to attenuate the first harmonic of the 60 Hz line noise, we used an elliptic IIR notch filter prior to extracting the broadband gamma band using bandpass filtering. The resulting signals were segmented into 50 ms windows with a 10 ms frameshift to capture the complex dynamics of neural activity underlying speech production. For each of these windows, we computed the signal energy and applied the natural logarithm to Gaussianize the data distribution.

We used context stacking to integrate information about temporal changes in the neural dynamics by augmenting each window with 8 neighboring windows ranging from -200 ms to +200 ms of neural activities. This results in feature vectors with $9 \cdot 64$ components.

The time-aligned acoustic speech data was transformed to logMel spectral features using the following steps: We first downsampled the acoustic data to 16 kHz and applied a segmentation with the same window size and frameshift as for the ECoG signals (50 ms and 10 ms, respectively). This preprocessing procedure allows to align ECoG features with spectral acoustic features while capturing speech-relevant information. In order to reduce the number of spectral features, we used triangular mel filter banks to finally extract 40 logarithmic mel-scaled spectral coefficients.

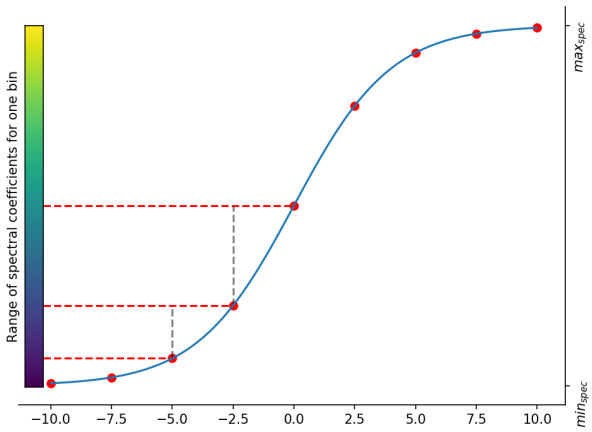

Figure 1: *Sigmoid quantization: Interval boundaries are defined using a sigmoid function. Given a uniform spacing (X-axis), a fixed number of intervals is defined with higher resolution toward the low and high ends of the spectral range. The colorbar applies to the range of one spectral bin of the logMel spectrogram.*

2.3. Quantization Approaches

In this study, we aim to transform the regression problem into a classification task. For this purpose, we convert the continuous feature space of the independent variable (spectral coefficients) into a discrete space by finding appropriate quantization intervals. Here, we focus on two approaches for quantization, i.e. (i) the *median-cut quantization* and (ii) the *sigmoid quantiziation* method. Both methods are based on a predefined number of quantization intervals. The first approach, a median-cut quantization [18, 19], is an algorithm that determines interval boundaries based on the number of occurrences of the spectral coefficients in the underlying data. In contrast, the second approach uses a sigmoid function with uniform spacing to define the interval boundaries. The sigmoid quantization method overcomes the imbalanced distribution of spectral coefficients, which typically occurs in speech data, where the amount of silence frames by far outnumber the speech frames.

The median-cut quantization is an iterative algorithm that works in the following way: Starting with all observed coefficients of a spectral bin, the algorithm splits the largest interval at its median into two intervals and iterates this splitting routine until a predefined amount of intervals is reached. Each of the resulting intervals is represented by one quantization value and all coefficients can be assigned based on the interval boundaries. In order to dequantize the spectrogram, each label gets replaced by the median value corresponding to its interval.

The sigmoid quantization relies on the logistic formula outlined in equation (1) to determine the boundaries of each interval.

$$f(x) = \frac{|min_{spec}| - |max_{spec}|}{1 + e^{-k \cdot x}} - |min_{spec}|, \qquad (1)$$

where min_{spec} and max_{spec} represent the minimum and maximum spectral coefficient and k defines the growth rate of the function. We used a uniform distribution for x to extract the interval boundaries based on the curve values. Figure 1 illustrates this procedure for the case of 8 intervals with a growth rate of $k = 0.5$. By using a logistic function, the granularity of the interval distribution can be tuned individually with the growth rate k, e.g. for coefficients in the low-frequency range such as silence as well as for coefficients in higher, speech-related frequencies.

2.4. Spectrogram and Waveform Reconstruction

For each logMel frequency bin, we use a linear discriminant analysis to predict the quantized spectral coefficient from the ECoG features. Due to the high number of ECoG features resulting from the number of electrodes multiplied by the temporal context, we performed a feature selection procedure prior to model training. For this purpose, the Spearman correlation coefficients between each ECoG feature and the mean spectral coefficients were calculated, ranked according to the highest correlation, and the top-150 features were selected.

After feature selection, the logMel spectral coefficients were discretized using the two outlined quantization approaches. With this quantization step, the regression problem was transformed to a classification task. Thus, a linear discriminant analysis could be used to find the maximum discriminability among the intervals. Although it is believed that the mapping between brain activity and speech outcome in terms of an acoustic signal is not linear [20], we focus here on a straightforward linear classifier. The initial experimental results presented in this paper are meant as a proof-of-concept to demonstrate the suitability of a discretized intermediate representation for speech decoding in general. We are convinced that more complex models, like deep neural networks will further improve the achieved performances, which we plan to investigate in the future.

For the transformation of estimated logMel spectral coefficients into an acoustic speech signal, we used the Griffin-Lim algorithm. In this iterative procedure, we limited ourselves to a number of 8 iterations in order to approximate the phase spectrogram and to obtain a time signal via the inverse Fourier transform.

3. Experimental Results

Experiments are performed in two steps: first, the best-case quantization error is estimated for a limited number of intervals assuming perfect performance. Second, decoding results based on linear models are compared with chance level estimated by breaking the temporal alignment of the data. These comparisons are carried out for both quantiziation approaches.

3.1. Quantization Error

We computed the quantization error over the number of intervals for both quantization methods to estimate the impact of discretization and to obtain the best-case decoding quality assuming perfect accuracy. Figure 2 displays the decrease of quantization error measured by the root mean squared error (RMSE) over the number of intervals, ranging from 3 to 22. As the number of intervals increase, the quantization error decreases and approaches 0 when the number of intervals approaches the resolution of the original feature space.

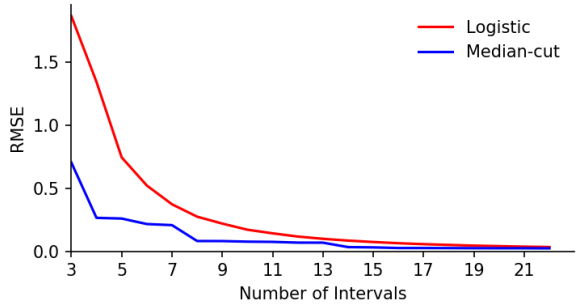

Figure 2: *Quantization error [in RMSE] over the number of intervals. RMSE measures the difference between reference and dequantized logMel spectral coefficients.*

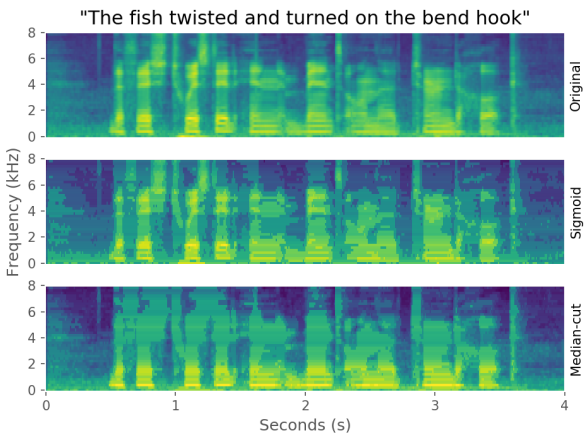

Figure 3: *logMel spectral coefficients prior to quantization (top) and after dequantization for sigmoid (middle) and median-cut (bottom) quantization approach assuming perfect accuracy. Both quantization approaches use 9 intervals.*

Figure 3 shows the logMel spectral coefficients of original speech (top) for one utterance along with the dequantized logMel spectra with respect to the sigmoid (middle) and median-cut (bottom) quantization approach, both using 9 intervals. This example indicates that the sigmoid quantization preserves more information in the higher frequencies while the median-cut quantization puts emphasis on silence portions and thus does not maintain a sufficiently fine resolution across the frequencies representing speech vocalization.

3.2. Spectrogram and Waveform Estimation

Speech decoding evaluation was performed using a 5-fold cross-validation. Linear models were trained on 80% of the data (40 utterances) and tested on the remaining 20% (10 utterances). Feature selection was performed in each fold on the training data only.

For evaluation, the Pearson correlation between original and estimated speech spectrograms was calculated by averaging over the correlation coefficients for each logMel frequency

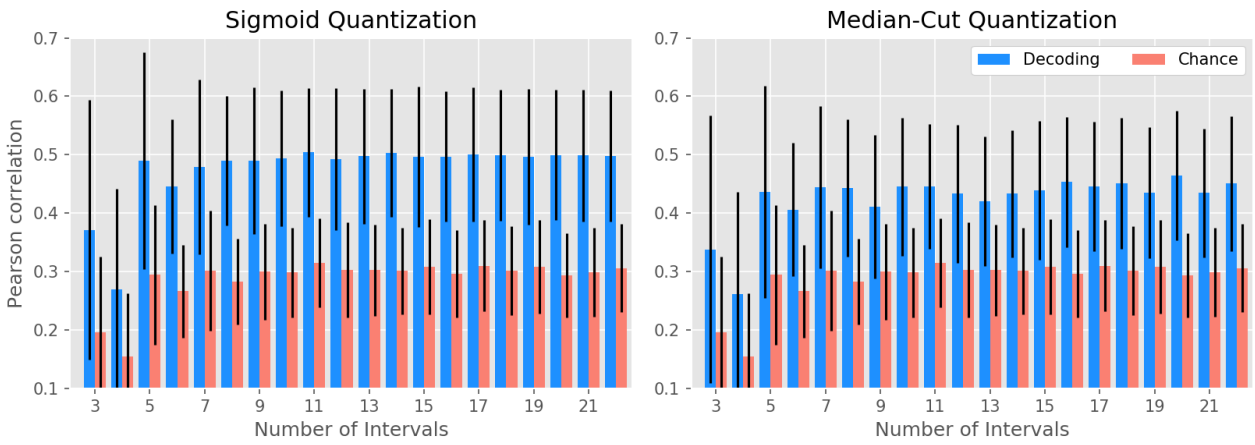

Figure 4: *Decoding results based on linear classification for sigmoid and median-cut quantization. Mean Pearson correlation (blue bars) and chance level (red bars) over number of intervals. Whiskers show standard deviation in 5-fold cross-validation.*

bin. Figure 4 shows the Pearson correlation coefficients between original and decoded spectrograms over the number of intervals. For the sigmoid quantization, results imply that a small number of quantization intervals (e.g. 5 or 7) perform as well as a larger number (blue bars). We also estimated a chance level (red bars) which breaks the alignment of the time-aligned simultaneous recordings of ECoG and logMel spectral coefficients by choosing a random point in time, splitting both data streams into two partitions at that point and swap the order of only one of the streams.

Reconstructed waveforms achieve an average objective intelligibility score [21] of 0.22 ± 0.02 and 0.27 ± 0.01 respectively, for the sigmoid and median-cut quantization approach across all intervals.

4. Discussion & Conclusion

The paper describes a proof-of-concept study towards synthesizing audible speech from brain activity data measured directly from the cortex based on a fixed set of discretized units. For the discretization of an intermediate representation, we investigated and compared two different approaches, the median-cut and the sigmoid quantization. Both methods are designed to counter the unbounded behavior of regression analyses to avoid potential biases caused by spurious amplitude spikes in neural data, while keeping the quantization error adjustable to preserve the desired spectral characteristics. The present results are limited to a single subject and further research is needed to conclude whether the results will generalize across subjects. In addition, further analysis is needed to determine how spurious amplitude spikes in the non-quantized condition compare to reconstructed speech in the quantized condition. Both limitations constitute important research questions which will provide more insights on how quantization methods can contribute to the challenging task of generating acoustic speech from brain activity data.

The experimental results in our study indicate that quantization is a feasible approach to reduce the spectral feature space to a small number of intervals while maintaining a close resemblance to the original speech signal. These outcomes based on straight-forward linear models achieved encouraging results indicating the applicability for speech synthesis from neural signals.

5. Acknowledgement

This work was supported by the National Science Foundation (1608140), USA and the Federal Ministry of Education and Research, Germany through the US-German Collaboration on Computational Neuroscience (project-ID 01GQ1602).

6. References

[1] T. Schultz, M. Wand, T. Hueber, K. D. J., C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 12, pp. 2257–2271, nov 2017.

[2] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[3] J. J. Shih and D. J. Krusienski, "Signals from intraventricular depth electrodes can control a brain–computer interface," *Journal of neuroscience methods*, vol. 203, no. 2, pp. 311–314, 2012.

[4] V. Guy, M.-H. Soriani, M. Bruno, T. Papadopoulo, C. Desnuelle, and M. Clerc, "Brain computer interface with the p300 speller: usability for disabled people with amyotrophic lateral sclerosis," *Annals of physical and rehabilitation medicine*, vol. 61, no. 1, pp. 5–11, 2018.

[5] M. Marchetti and K. Priftis, "Effectiveness of the p3-speller in brain–computer interfaces for amyotrophic lateral sclerosis patients: a systematic review and meta-analysis," *Frontiers in neuroengineering*, vol. 7, p. 12, 2014.

[6] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ecog: A pilot study," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 1540–1543.

[7] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[8] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[9] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, p. 036019, 2019.

[10] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[11] C. Herff, L. Diener, M. Angrick, E. M. Mugler, M. C. Tate, M. Goldrick, D. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor and inferior frontal cortices," *Frontiers in Neuroscience*, vol. 13, p. 1267, 2019.

[12] M. Angrick, C. Herff, G. Johnson, J. Shih, D. Krusienski, and T. Schultz, "Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings," *Neurocomputing*, vol. 342, pp. 145–151, 2019.

[13] IEEE, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[14] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[15] E. Leuthardt, X.-M. Pei, J. Breshears, C. Gaona, M. Sharma, Z. Freudenburg, D. Barbour, and G. Schalk, "Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task," *Frontiers in human neuroscience*, vol. 6, p. 99, 2012.

[16] N. Crone, L. Hao, J. Hart, D. Boatman, R. P. Lesser, R. Irizarry, and B. Gordon, "Electrocorticographic gamma activity during word production in spoken and sign language," *Neurology*, vol. 57, no. 11, pp. 2045–2053, 2001.

[17] V. L. Towle, H.-A. Yoon, M. Castelle, J. C. Edgar, N. M. Biassou, D. M. Frim, J.-P. Spire, and M. H. Kohrman, "Ecog gamma activity during a language task: differentiating expressive and receptive speech areas," *Brain*, vol. 131, no. 8, pp. 2013–2027, 2008.

[18] P. Heckbert, "Color image quantization for frame buffer display," *ACM Siggraph Computer Graphics*, vol. 16, no. 3, pp. 297–307, 1982.

[19] L. Diener, T. Umesh, and T. Schultz, "Improving fundamental frequency generation in emg-to-speech conversion using a quantization approach," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 682–689.

[20] S. Martin, J. d. R. Millán, R. T. Knight, and B. N. Pasley, "The use of intracranial recordings to decode human language: Challenges and opportunities," *Brain and language*, vol. 193, pp. 73–83, 2019.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.