DRUM TRANSCRIPTION FROM MULTICHANNEL RECORDINGS WITH NON-NEGATIVE MATRIX FACTORIZATION

David S. Alves[1,2], *Jouni Paulus*[2], *and José Fonseca*[1]

[1]Department of Electrical Engineering
Faculty of Science and Technology
New University of Lisbon, Portugal
email:davsantosster@gmail.com

[2]Department of Signal Processing
Tampere University of Technology
Tampere, Finland
email:jouni.paulus@tut.fi

ABSTRACT

Automatic drum transcription enables handling symbolic data instead of plain acoustic information in music information retrieval applications. Usually the input to the transcription system is single-channel audio, and as a result the proposed solutions are designed for this kind of input. However, in studio environment the multichannel recording of the drums is often available. This paper proposes an extension to a non-negative matrix factorization drum transcription method to handle multichannel data. The method creates spectral templates for all target drums from all available channels, and in transcription estimates time-varying gains for each of them so that the sum approximates the recorded signal. Sound event onsets are detected from the estimated gains. The system is evaluated with multichannel data from a publicly available data set, and compared with other methods. The results suggest that the use of multiple channels instead of a single-channel mix improves the transcription result.

1. INTRODUCTION

Drum transcription provides a method to transition from acoustic signal to a symbolic representation of the drum sound content. Drum transcription means the process of detecting the occurrences of drum sound events in a musical performance and recognizing the instruments used. The input performance is usually a single-channel or a stereo recording. Several methods for solving the transcription from material containing only drum hits or from polyphonic music have been proposed in literature, see [5] for a review. The methods are often divided into two categories: event-based and source separation -based methods. The method in the first category locate sound events in the input signal and then recognize the content. In the source separation approaches, the individual target drums are considered to be the sound sources that are separated from the mixture signal, and the sound event onsets are searched from the separated drum signals. The method proposed in this paper belongs to the separation-based approaches. Other recent drum transcription methods include using spectrogram template matching and adaptation [14], using support vector machine classifiers to recognize the sound event content [7], and using continuous hidden Markov models to perform the signal segmentation and classification simultaneously [11].

The input signal is represented as a magnitude spectrogram $\boldsymbol{X}$ with F rows corresponding to frequencies and T columns corresponding to frames. In this representation, the mixture signal can be considered to be a sum of the spectrograms of the N source signals

$$\boldsymbol{X} = \sum_{n=1}^{N} \boldsymbol{X}_n + \varepsilon, \tag{1}$$

where $\boldsymbol{X}_n$ is the spectrogram of the nth source, and ε represents the approximation error. Each of the sources $\boldsymbol{X}_n$ is considered to be a product of two basis vectors

$$\boldsymbol{X}_n = \boldsymbol{s}_n \boldsymbol{a}_n^T. \tag{2}$$

With this approximation and omitting the approximation error term, (1) can be rewritten as a matrix product

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}, \tag{3}$$

where the component matrices $\boldsymbol{S}$ and $\boldsymbol{A}$ are defined as

$$\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N] \tag{4}$$

and

$$\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_N]^T. \tag{5}$$

Several approaches have been proposed for solving the decomposition of $\boldsymbol{X}$ into the components $\boldsymbol{S}$ and $\boldsymbol{A}$. One of the first was independent subspace analysis (ISA) [1]. ISA was proposed to solve the problem of applying independent component analysis on a single-channel signal by representing the signal as a spectrogram and considering each frequency band as a channel signal. In this context, each vector $\boldsymbol{s}_n$ can be considered as a frequency basis function and $\boldsymbol{a}_n$ the corresponding time basis function.

ISA performs the separation without any prior knowledge of the sources despite that in many cases some information is available. In [4] prior subspace analysis (PSA) was proposed to utilize prior knowledge of the sources by calculating spectral templates $\boldsymbol{s}_n$ for each target drum in advance and then calculating the corresponding time-varying gains $\boldsymbol{a}_n$. The sound events are then detected from the temporal basis functions.

Another way to perform the decomposition of (3) is to use non-negative matrix factorization (NMF) [13], which assumes that every element in the matrices $\boldsymbol{X}$, $\boldsymbol{A}$, and $\boldsymbol{S}$ are non-negative. This non-negativity constraint fits the magnitude spectrogram representation as it has only non-negative values. Similar to the idea of PSA, use of spectral templates with NMF for drum transcription was proposed in [12]. This paper extends that work.

When a drum kit is recorded in a studio environment, usually each membranophone (drums with membranes, e.g.,

This work was supported by the Academy of Finland, project no. 5213462 (Finnish Centre of Excellence program 2006 - 2011).

kick drum, snare drum, tom-toms) has at least one close microphone and all the cymbals are recorded with two or more overhead microphones. The knowledge of the content and timing of each hit in the recorded signals would enable more flexible editing of the recordings. Since these multichannel signals are available, the transcription is done using them instead of the monophonic or stereophonic mix-down available in the later stages of record production.

The rest of this paper is organized as follows. Section 2 describes the use of NMF for drum transcription starting from the single-channel method and then extending it for multichannel data. Section 3 details the evaluations of the system performance, including the material used, the performance metrics, and the evaluation results. Finally, Section 4 provides the conclusions of the paper.

2. DRUM TRANSCRIPTION USING NON-NEGATIVE MATRIX FACTORIZATION

The factorization of non-negative matrix $\boldsymbol{X}$ to two non-negative matrices $\boldsymbol{S}$ and $\boldsymbol{A}$ can be done using the Lee and Seung algorithm [10] which minimizes the Kullback-Leibler divergence -like distance measure

$$D(\boldsymbol{X}||\boldsymbol{SA}) = \sum_{f,t} [\boldsymbol{X}]_{f,t} \log \frac{[\boldsymbol{X}]_{f,t}}{[\boldsymbol{SA}]_{f,t}} - [\boldsymbol{X}]_{f,t} + [\boldsymbol{SA}]_{f,t} \qquad (6)$$

between the input $\boldsymbol{X}$ and the approximation $\boldsymbol{SA}$. The matrices $\boldsymbol{S}$ and $\boldsymbol{A}$ are initialized with non-negative random values and then iteratively updated with the multiplicative rules

$$\boldsymbol{A} \leftarrow \boldsymbol{A}.* \frac{\boldsymbol{S}^T(\boldsymbol{X}./(\boldsymbol{SA}))}{\boldsymbol{S}^T\boldsymbol{1}} \qquad (7)$$

and

$$\boldsymbol{S} \leftarrow \boldsymbol{S}.* \frac{(\boldsymbol{X}./(\boldsymbol{SA}))\boldsymbol{A}^T}{\boldsymbol{1}\boldsymbol{A}^T}. \qquad (8)$$

In the equations above, $\boldsymbol{1}$ is a all-one matrix of size $F \times T$. Element wise multiplication and division operations are represented by $.*$ and $./$ after [8]. For a tutorial-like overview of other NMF algorithms and the relationship of NMF with its generalization non-negative tensor factorization (NTF), refer to [2]. Using NTF for sound source separation is demonstrated, e.g., in [3].

2.1 Single Channel Input Data

The use of NMF and templates in drum transcription was proposed in [12]. The spectral basis functions (or templates) $\boldsymbol{s}_n$ are calculated from signals containing only the target drum. This signal is factorized into one source and the resulting $\boldsymbol{S}$ containing only one column is assumed to represent the main characteristics of the target drum, and it is taken as the spectral template for that drum. If several training signals are available for one drum, the spectral template is calculated separately for each of them and then averaged. The templates of all target drums are combined with (4) to create the spectral template matrix $\boldsymbol{S}$.

The time-varying gains are obtained from the spectrogram of the input signal by applying only the update formula (7) with the spectral templates until the result has converged. The detection of the drum hits from the resulting time-varying gains $\boldsymbol{a}_n$ is done by applying a psychoacoustically motivated onset detection on the gain curves. The process consists of μ-law compression, low-pass filtering, differentiation, half-wave rectification, detecting peaks, and thresholding the located peaks. The optimal threshold value is determined for each target drum using a set of training signals by minimizing the sum of extraneous and missed detections on them. The onset detection method is motivated by [9].

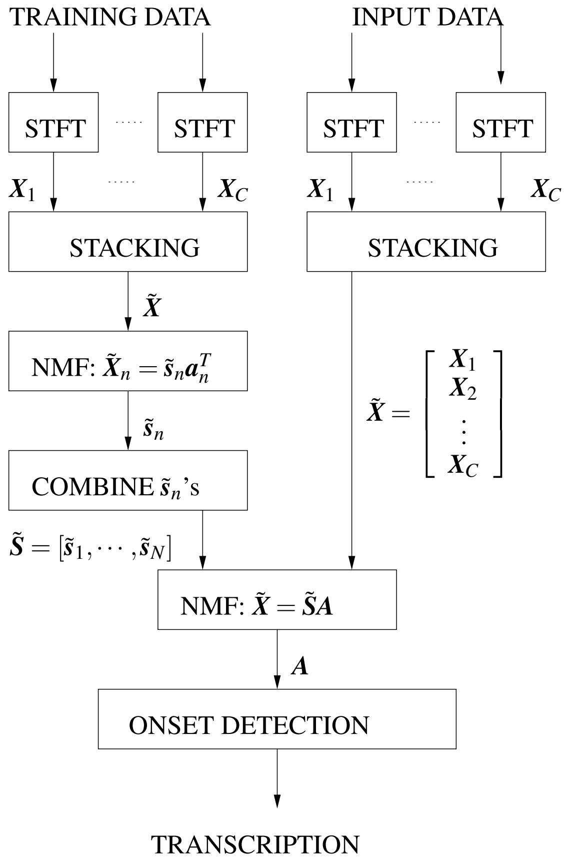
TRAINING DATA
INPUT DATA
STFT
STFT
STFT
STFT
X_1
X_C
X_1
X_C
STACKING
STACKING
$\tilde{X}$
NMF: $\tilde{X}_n = \tilde{s}_n a_n^T$
$\tilde{s}_n$
COMBINE $\tilde{s}_n$'s
$\tilde{S} = [\tilde{s}_1, \cdots, \tilde{s}_N]$
$\tilde{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_C \end{bmatrix}$
NMF: $\tilde{X} = \tilde{S}A$
A
ONSET DETECTION
TRANSCRIPTION

Figure 1: Functional overview of the proposed method. It calculates the spectral template $\tilde{\boldsymbol{s}}_n$ over all channels from signals containing only the target drum. The process is repeated for all target drums, and the resulting templates are combined to create the spectral template matrix $\tilde{\boldsymbol{S}}$. The templates are used to estimate the time-varying gains of the target drums played in the multichannel input. Finally, the sound event onsets are detected from the gains.

The single-channel NMF method has proven to perform well when the input data match the model well, i.e., there are only target drums in the mixture [12]. Additional drums and other sound sources caused the performance to degrade quickly. Furthermore, if the target drums overlap considerably in the frequency domain, the factorization may produce an undesired result.

2.2 Multichannel Input Data

We propose extending the single-channel method so that the factorized spectrogram is calculated not only from a single-channel mixture, but from the multiple microphone signals available in the studio environment. The spectrograms of C individual channels $\boldsymbol{X}_c$, $c \in 1 \ldots C$ are stacked to form the matrix to be factorized

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_C \end{bmatrix}. \tag{9}$$

Now the factorization (3) is extended to

$$\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{S}}\boldsymbol{A}, \tag{10}$$

where the spectral templates are similarly stacked into

$$\tilde{\boldsymbol{S}} = \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_2 \\ \vdots \\ \boldsymbol{S}_C \end{bmatrix} = \begin{bmatrix} \boldsymbol{s}_{1,1}, \boldsymbol{s}_{1,2}, \cdots, \boldsymbol{s}_{1,N} \\ \boldsymbol{s}_{2,1}, \boldsymbol{s}_{2,2}, \cdots, \boldsymbol{s}_{2,N} \\ \vdots \cdots \ddots \vdots \\ \boldsymbol{s}_{C,1}, \boldsymbol{s}_{C,2}, \cdots, \boldsymbol{s}_{C,N} \end{bmatrix} = [\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, \cdots, \tilde{\boldsymbol{s}}_N]. \tag{11}$$

In the matrix above, $\boldsymbol{s}_{c,n}$ is the spectral template of nth source in the microphone channel c. The main motivation for this stacking is that it allows to separate drums that have similar spectral content, e.g., tom-toms, if their intensity in different microphone channels differs. A functional overview of the proposed method is presented in Figure 1.

The training and use of the model is similar to the one of the single-channel method with small exceptions. The main difference is that the template $\tilde{\boldsymbol{s}}_n$ has to be calculated over all channels at once instead of constructing it by stacking the templates from individual channels. This is required to ensure that the relative level differences between the channels would be correct in the resulting template. The use of the template in the factorization and the determination of the onset instants is similar to the single-channel case.

3. EVALUATIONS

The performance of the proposed method was tested using a data set of multichannel drum recordings with manually made annotations [6]. The performance is compared with two other methods.

3.1 Acoustic Material

The material used was from the public audio subset of "ENST drums" database [6]. The database contains recordings for three different drummers and drum kits. Each drummer has a different number of instruments and a different microphone setup. The material consists of individual drum hits, traditional short drum sequences, and drum tracks played along accompaniment that are provided separately from the drum audio. These "minus one" tracks are further divided into tracks with acoustic and with MIDI accompaniment. For all of the material, the database contains individual microphone signals (7 or 8, depending on the kit), a "dry" mix-down with only level adjustments, and a "wet" mix-down with compressor and other effects. The evaluations use the individual drum hits for template calculation, and the "minus one" tracks for testing. In total, there are 64 "minus one" tracks in the data set with duration ranging from 30 s to 75 s and mean duration of 55 s.

3.2 Evaluation Process

The target drum set in the evaluations consists of bass drum (BD), snare drum (SD), and hi-hat (HH). Other drums in the tracks are not attempted to be transcribed. The target drum set is limited in this way since the three drums for the main rhythmic background on a large body of the pop/rock songs, while the other drums provide mainly accentuations. The input to the system consists of the close microphone signals and contains only drums sounds. The evaluations are done using leave-one-out cross-validation scheme on each of the three drum kits separately, and the presented results are calculated over all cross-validation folds. The cross-validation has to be made for each drummer subset separately because the method relies on the channel setup to remain identical across training and testing phases.

The spectrogram signal representation uses the frequency resolution of 24 Bark bands, and the analysis window length is 24 ms with 75% overlap.

3.3 Performance Metrics

The hits in the obtained result and in the ground truth are matched. Two hits were accepted as a match if they differ less than 50 ms in their onset times. Recall rate R_R is the ratio of correct hits to the hits in the ground truth, precision rate R_P the ratio of correct hits to the hits in the result, and F-measure is the harmonic mean of these $F = 2R_R R_P/(R_R + R_P)$.

3.4 Comparison to Other Systems

The performance of the proposed system is compared to other methods. First, the proposed multichannel version is compared with the single-channel NMF transcription method [12]. The single-channel data used in the experiments are the "dry-mix" versions. This comparison allows to determine if the use of multiple channels provides any additional information that the method can use compared to the mix-down.

Secondly, a naive multichannel transcription method is implemented. It assumes that each close microphone channel contains mostly the sound of that drum. The transcription can then be made by detecting sound event onsets on the channel signal. The onset detection is made with the method proposed in [9], and the detection threshold is determined using training data in a manner similar to the proposed method. The main weakness of this naive approach is that it relies on every target drum to have a close microphone.

Finally, a second single-channel comparison method is the one proposed in [7] attempting to enhance the drum sound content when there is also accompaniment involved. Even though the system was designed to transcribe drums with accompaniment, the original publication also provides evaluation results in the case when the accompaniment is not present. The evaluations in the original publication use the same data set and evaluation metric as this paper. These results are gathered in Table 2. When comparing the result, it should be remembered that the results presented in [7] were obtained using a cross-validation scheme differing from the one we are using, and used only single-channel input.

3.5 Results and Discussion

The evaluation results are presented in Tables 1-3. Table 3 presents the detailed results for the three different drum kit subsets for all of the "minus one" tracks. The results in Table 1 show that the naive approach works surprisingly well on the multichannel recordings. However, the multichannel approach performs even better. Both the single-channel

		BD	SD	HH	Total
1-channel [12]	R_R	93.7%	51.4%	77.1%	74.0%
	R_P	95.2%	80.0%	67.8%	78.5%
	F	94.4%	62.6%	72.1%	76.2%
onsets	R_R	85.7%	54.4%	83.5%	75.3%
	R_P	85.5%	76.8%	76.7%	79.4%
	F	85.6%	63.6%	80.0%	77.3%
m-channel	R_R	95.9%	77.8%	87.9%	87.1%
	R_P	93.5%	77.5%	78.7%	82.5%
	F	94.7%	77.6%	83.0%	84.7%

Table 1: Evaluation results for the proposed method (m-channel), the single-channel NMF (1-channel), and the naive method (onsets), on the "minus one" tracks of the ENST public dataset.

and naive multichannel method have a relatively low recall rate on snare drum. We assume that this may be caused by the "ghost hits", which are very light hits producing more full-bodied rhythmic feel. The lightness of these hits causes difficulties in setting the onset detection threshold. The difference in the recall and precision rates for hi-hat may be partially caused by the presence of the other cymbals in the tracks: since they do not belong to the target set and their spectral properties overlap with hi-hat, their presence causes some extra hits to be detected. More detailed inspection on the material revealed that a large body of the hi-hat insertions were caused by the cow bell instrument played in latin tracks in the place of hi-hat.

Based on the total F-measure of each individual target piece, the performance difference between the single-channel NMF and naive onset detection based method is not statistically significant. However, the overall performance increase obtained with the multichannel NMF method over the comparison methods is statistically significant with the level $p > 99\%$.

When the evaluations are done on the subset of "minus one" tracks played on a real accompaniment, the performance of both single-channel and multichannel NMF decrease, as can be seen in Table 2. Only the bass drum performance remains high. Comparing to the reference method [7], the proposed system is more accurate on bass drum, slightly more accurate on snare drum, and hi-hat performance is approximately even. Some of the performance degradation may be caused by the small amount of training data (only nine tracks for each drummer).

The per-drummer results in Table 3 show that there are some performance difference between the three subsets. The main improvement of the proposed method is visible in the snare drum results, where both recall and precision rates are greatly improved.

4. CONCLUSION

This paper has presented an extension of an NMF based drum transcription method to multichannel data. The multichannel data are available from the recording setup in a studio environment. The proposed method creates spectral templates for each target drum to each input channel, and uses NMF to recover the time-varying gains for the drums. Finally, onsets are searched from the recovered gains. The proposed system has been evaluated on multichannel data and compared to other methods in the task of transcribing bass drum, snare drum, and hi-hat. The results show that the use of multiple channels increases the system performance over the single-channel method.

		BD	SD	HH
reference [7]	R_R	70.0%	64.2%	86.5%
	R_P	79.8%	71.0%	73.6%
	F	74.6%	67.4%	79.5%
1-channel [12]	R_R	94.1%	47.9%	70.6%
	R_P	95.1%	71.0%	63.4%
	F	94.6%	57.2%	66.8%
m-channel	R_R	96.0%	72.0%	84.0%
	R_P	93.6%	74.3%	75.5%
	F	95.0%	73.0%	79.4%

Table 2: Evaluation results for the proposed method, the single-channel NMF, and the reference method, on the "minus one" tracks excluding the ones recorded with MIDI accompaniment (the results for the reference are from [7]).

REFERENCES

[1] M. A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proc. of International Computer Music Conference*, pages 154–161, Berlin, Aug. 2000.

[2] A. Cichocki, R. Zdunek, and S. Amari. Nonnegative matrix and tensor factorization. *IEEE Signal Processing Magazine*, pages 142–145, Jan. 2008.

[3] D. FitzGerald, M. Cranitch, and E. Coyle. Sound source separation using shifted non-negative tensor factorisation. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006.

[4] D. FitzGerald, B. Lawlor, and E. Coyle. Prior subspace analysis for drum transcription. In *Proc. of 114th Audio Engineering Society Convention*, Amsterdam, Mar. 2003.

[5] D. FitzGerald and J. Paulus. Unpitched percussion transcription. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*, pages 131–162. Springer, 2006.

[6] O. Gillet and G. Richard. ENST-Drums: an extensive audio-visual database for drum signal processing. In *Proc. of 7th International Conference on Music Information Retrieval*, Victoria, B.C., Canada, Oct. 2006.

[7] O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):529–540, Mar. 2008.

[8] M. Helén and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. of 13th European Signal Processing Conference*, Antalya, Turkey, Sept. 2005.

[9] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Ariz., USA, 1999.

[10] D. D. Lee and H. S. Seung. Algorithms for non-

		BD			SD			HH			Total		
		D1	D2	D3	D1	D2	D3	D1	D2	D3	D1	D2	D3
1-channel [12]	R_R	87.0%	97.5%	97.5%	47.7%	65.0%	41.4%	67.1%	80.4%	82.0%	67.1%	80.0%	74.9%
	R_P	94.6%	95.7%	95.2%	76.7%	82.3%	80.7%	55.9%	71.5%	74.6%	72.4%	80.5%	82.3%
	F	90.6%	96.6%	96.4%	58.8%	72.6%	54.7%	61.0%	75.7%	78.2%	69.7%	80.3%	78.4%
onsets	R_R	83.0%	86.4%	88.0%	73.5%	44.9%	43.8%	75.9%	82.8%	90.6%	77.4%	72.0%	76.4%
	R_P	75.6%	94.6%	89.7%	74.3%	73.8%	85.3%	77.0%	73.4%	79.7%	75.7%	78.9%	83.7%
	F	79.1%	90.3%	88.8%	73.9%	55.8%	57.9%	76.4%	77.8%	84.8%	76.5%	75.3%	79.9%
m-channel	R_R	93.2%	97.4%	97.4%	80.0%	77.8%	75.4%	86.0%	86.9%	90.4%	86.4%	86.8%	88.1%
	R_P	93.8%	95.3%	91.7%	77.4%	82.4%	73.0%	76.6%	77.1%	82.1%	81.9%	83.2%	82.3%
	F	93.5%	96.4%	94.5%	78.7%	80.1%	74.2%	81.1%	81.7%	86.0%	84.1%	85.0%	85.1%

Table 3: Evaluation results for the individual drummers (different drummers are denoted with D1, D2, and D3) for the same methods and material as presented in Table 1.

negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 556–562. MIT Press, 2001.

[11] J. Paulus and A. Klapuri. Combining temporal and spectral features in HMM-based drum transcription. In *Proc. of 8th International Conference on Music Information Retrieval*, pages 225–228, Vienna, Sept. 2007.

[12] J. Paulus and T. Virtanen. Drum transcription with non-negative spectrogram factorisation. In *Proc. of 13th European Signal Processing Conference*, Antalya, Turkey, Sept. 2005.

[13] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. of 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, New Platz, N.Y., USA, Oct. 2003.

[14] K. Yoshii, M. Goto, and H. G. Okuno. Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):333–345, Jan. 2007.