Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition

Antti Hurmalainen[1], *Rahim Saeidi*[2], *Tuomas Virtanen*[1]

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
antti.hurmalainen@tut.fi, rahim.saeidi@let.ru.nl, tuomas.virtanen@tut.fi

Abstract

Spectrogram factorisation using a dictionary of spectro-temporal atoms has been successfully employed to separate a mixed audio signal into its source components. When atoms from multiple sources are included in a combined dictionary, the relative weights of activated atoms reveal likely sources as well as the content of each source. Enforcing sparsity on the activation weights produces solutions, where only a small number of atoms are active at a time. In this paper we propose using group sparsity to restrict simultaneous activation of sources, allowing us to discover the identity of an unknown speaker from multiple candidates, and further to recognise the phonetic content more reliably with a narrowed down subset of atoms belonging to the most likely speakers. An evaluation on the CHiME corpus shows that the use of group sparsity improves the results of noise robust speaker identification and speech recognition using speaker-dependent models.

Index Terms: group sparsity, speech recognition, speaker identification, spectrogram factorization

1. Introduction

In several studies it has been reported, how spectrogram factorisation using a dictionary of atoms has produced strong results in separating multiple non-stationary sources from mixed observations [1, 2, 3]. However, a common assumption is that only certain sources are active in the mixture — for example, one known speaker over background noise, or two known speakers. Under this assumption, only the relevant dictionaries are chosen for the factorisation task, thus reducing problem complexity and confusion with sources not present in the mixture. In reality, the set of potential sources may be significantly larger than the number of sources active in the mixture, and the identities of active sources may not be known beforehand. There is ongoing research on multi-talker tasks with non-negative matrix factorisation (NMF) given as one option, but thus far the performance of its basic form has not been found satisfactory [4].

It has been shown that activations of dictionary atoms acquired via NMF can act as evidence for both the speaker identity and the phonetic content of speech [3, 5, 6]. Enforcing sparsity on the activations improves the classification results [5]. Therefore the method is referred to as *sparse classification* (SC). A straightforward sparsity constraint is to penalise all non-zero activation weights by adding a weighted L_1 norm of all activations to the cost function to be minimised. The problem of this approach is that the acquired solution may contain atoms from any number of sources as long as the distribution of individual atoms is sparse. The same spectral features may carry a different meaning if taken from another source, thereby harming the classification outcome. If we expect only a limited number of sources to be active at a time, it would be beneficial to exploit this knowledge by enforcing corresponding structure on the activations, that is, to prefer solutions where activations appear as groups matching to a few sources at a time.

Group sparsity allows defining groups of dictionary atoms and constraining the factorisation to use only a small number of groups with active atoms. The technique has been previously employed in some applications, including image classification [7], music separation [8], DNA sequences [9], and automatic speech recognition [10]. In this paper we propose using group sparsity in addition to common L_1 sparsity to produce factorisation solutions, where a narrowed down set of speakers is active at a time. Furthermore, we propose an algorithm which favours the same speakers over the whole duration of an utterance. Sparse activations are shown to produce improved speaker and speech recognition results in a task, where an utterance from an unknown speaker must be recognised among additive noise.

The paper is organised as follows. Section 2 describes the core concepts of spectrogram factorisation and sparse classification. In Section 3 we derive a model and a corresponding iterative update rule to induce consistent group sparsity in utterances comprising multiple observation windows. Experimental set-up on CHiME data is presented in Section 4. Results, discussion and conclusions follow in Sections 5, 6, and 7.

2. Non-negative spectrogram factorisation

Our separation framework is based on representing a mixed observation spectrogram as a linear, non-negative combination of *atoms* — spectrogram segments acquired from sources such as single speakers or background noise. Each atom is modelled with a $B \times T$ magnitude spectrogram matrix, where B is the number of frequency bands and T is *window length* — the number of consecutive frames in an atom. We model noisy speech with J speech and K noise atoms, together forming a *dictionary* (or *basis*) of $L = J + K$ atoms. If we reshape the atoms into length $B \cdot T$ vectors $\mathbf{a}_j^{\mathrm{s}}$ ($j \in [1, J]$) and $\mathbf{a}_k^{\mathrm{n}}$ ($k \in [1, K]$) for speech and noise, respectively, a similarly vectorised observation $\mathbf{y}$ can be estimated as a linear sum

$$\mathbf{y} \approx \sum_{j=1}^{J} \mathbf{a}_j^{\mathrm{s}} x_j^{\mathrm{s}} + \sum_{k=1}^{K} \mathbf{a}_k^{\mathrm{n}} x_k^{\mathrm{n}} \qquad (1)$$

where x_j^{s} and x_k^{n} are the *activation weights* of speech and noise atoms. The same equation can be given in a matrix form as

$$\mathbf{y} \approx \mathbf{A}^{\mathrm{s}}\mathbf{x}^{\mathrm{s}} + \mathbf{A}^{\mathrm{n}}\mathbf{x}^{\mathrm{n}} \qquad (2)$$

where the columns of matrices $\mathbf{A}^{\mathrm{s}}$ and $\mathbf{A}^{\mathrm{n}}$ consist of vectorised speech and noise atoms, and $\mathbf{x}^{\mathrm{s}}$ and $\mathbf{x}^{\mathrm{n}}$ are *activation vectors* for speech and noise, together denoted by vector $\mathbf{x}$ of length L.

In previous work, we have experimented with two different methods to model *observation spectrograms* $\mathbf{Y}$ ($B \times F$), where the number of frames F is larger than T [11]. The first uses $W = F - T + 1$ overlapping windows, each factorised independently. The second, convolutive model is similar but produces a joint spectrogram estimate $\mathbf{\Psi}$ from all window indices simultaneously. Both produce an $L \times W$ *activation matrix* $\mathbf{X}$, each of its columns containing an activation vector for a window index. The previously used cost function to be minimised consists of Kullback-Leibler divergence between the observation spectrogram $\mathbf{Y}$ and its estimate $\mathbf{\Psi}$

$$d_{\mathrm{KL}}(\mathbf{Y}, \mathbf{\Psi}) = \sum_{(y,\psi) \in (\mathbf{Y},\mathbf{\Psi})} y \log \frac{y}{\psi} - y + \psi \qquad (3)$$

and the L_1 norm of $\mathbf{X}$ multiplied elementwise by a sparsity penalty matrix $\mathbf{\Lambda}_1$,

$$f_1 = ||\mathbf{X} \otimes \mathbf{\Lambda}_1||_1. \qquad (4)$$

Iterative updates rules to find $\mathbf{X}$ for these costs and for both temporal models can be found in earlier work [3, 11]. In this work, we extend the convolutive model to support group sparsity in addition to basic L_1 sparsity. The same approach for group sparsity also applies to independent window factorisation.

3. Group sparsity for activation matrices

3.1. Multi-column matrix group sparsity

A generalised form of group sparsity can be achieved by using a cost function

$$f_{\mathrm{g}} = ||\sqrt{\mathbf{G}^2\mathbf{X}^2}||_1 \qquad (5)$$

on the activation matrix $\mathbf{X}$. Here $\mathbf{G}$ is a $S \times L$ matrix assigning the L atom indices to S groups with any weights. Square and square root operations are elementwise. The function measures weighted L_2 norms within groups for each window index, produces a $S \times W$ matrix of group 2-norms, and sums them over all groups and window indices. Because in this work we use group sparsity for selection of groups, that is, denoting basic membership without further atom weighting, we simplify the structure by limiting ourselves to assignment matrices of type $\mathbf{G} = \lambda_{\mathrm{g}}\mathbf{G}_{\mathrm{B}}$, where $\mathbf{G}_{\mathrm{B}}$ is a binary matrix denoting atom membership in groups, and λ_{g} is a common weight factor for all chosen atoms. The simplified cost for binary matrices is

$$f_{\mathrm{g}} = \lambda_{\mathrm{g}}||\sqrt{\mathbf{G}_{\mathrm{B}}\mathbf{X}^2}||_1. \qquad (6)$$

However, the given cost function measures group sparsity independently for each window. Although the columns of $\mathbf{X}$ each become sparse on a group level, they may all have different groups active. In our speech recognition task, we expect the same speaker to be active over all window indices within a short observation. Therefore we modify the function to measure the group L_2 norms for summed activity over window indices, $\mathbf{x}_\Sigma = \mathbf{X} \cdot \mathbf{1}$ ($\mathbf{1}$ being an all-one column vector of length W). The cost function becomes

$$f_{\mathrm{g}} = \lambda_{\mathrm{g}}||\sqrt{\mathbf{G}_{\mathrm{B}}\mathbf{x}_\Sigma^2}||_1. \qquad (7)$$

3.2. Combined group and atom sparsity

The equations given in Section 3.1 introduce sparsity over groups, but not over single atoms within a group. Because we have earlier found atom-level sparsity beneficial in SC-based speech recognition as well, both are combined for a cost function that induces sparsity over atoms, yet prefers solutions where the activations come from a sparse set of groups. The total cost function for KL-divergence, group sparsity and L_1 sparsity is

$$f_{\mathrm{tot}} = d_{\mathrm{KL}}(\mathbf{Y}, \mathbf{\Psi}) + \lambda_{\mathrm{g}}||\sqrt{\mathbf{G}_{\mathrm{B}}\mathbf{x}_\Sigma^2}||_1 + ||\mathbf{X} \otimes \mathbf{\Lambda}_1||_1. \qquad (8)$$

3.3. Iterative update algorithm

The total cost function (8) is minimised by initialising all the entries in the activation matrix $\mathbf{X}$ to unity, and then updating it iteratively with an update rule

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\sum_{t=1}^{T} \mathbf{A}_t^T \overset{\leftarrow(t-1)}{[\frac{\mathbf{Y}}{\mathbf{\Psi}}]}}{\sum_{t=1}^{T} \mathbf{A}_t^T \overset{\leftarrow(t-1)}{\mathbf{1}} + \mathbf{\Lambda}_{\mathrm{g}} + \mathbf{\Lambda}_1}. \qquad (9)$$

Here each $\mathbf{A}_t$ is a $B \times L$ matrix containing frame t of all basis atoms. Operator $\leftarrow$ shifts matrix columns left, followed by truncation to W columns. Estimated utterance spectrogram $\mathbf{\Psi}$ is calculated by

$$\mathbf{\Psi} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}}. \qquad (10)$$

with shifting right ($\rightarrow$) taking place in a $L \times F$ zero-padded matrix. Matrix $\mathbf{\Lambda}_{\mathrm{g}}$ defines the group sparsity cost of each atom and is updated within each iteration based on the activation sum vector. Its columns are identical and are given as

$$\boldsymbol{\lambda}_{\mathrm{g}} = \lambda_{\mathrm{g}}\mathbf{x}_\Sigma \otimes (\mathbf{G}_{\mathrm{B}}^T(\mathbf{G}_{\mathrm{B}}\mathbf{x}_\Sigma^2)^{-1/2}). \qquad (11)$$

4. Application to speaker identification and speech recognition

4.1. CHiME data and feature space

To study the potential of group sparsity in finding a sparse combination of sources, we ran experiments on CHiME data, containing GRID command utterances from 34 speakers over family household noises at SNRs ranging from +9 to -6 dB [12]. The utterances follow a linear *verb-colour-preposition-letter-digit-coda* grammar. A default language model utilising 250 sub-word states for the 51 word vocabulary is provided. The data consists of three sets:

- Train: 500 utterances from each speaker without additive noise ('clean')
- Development: a set of 600 utterances from all speakers combined, repeated over six SNRs
- Test: as development, but with different utterances and noise content

All audio data, including 'clean' sets, has room reverberation. 16 kHz binaural files were used for the experiments. All audio was converted into spectrogram features with $B = 40$ Mel scale spectral bands, 25 ms frame length, 10 ms frame shift, and averaging of the magnitude spectrograms of left and right channels. The bands were linearly scaled using a fixed scaling based on speech training data [3]. Atom length T was set to 25 frames (265 ms).

4.2. Bases and sparsity parameters

We created a 250-atom speech basis for each speaker by modelling the spectrogram context of each state in turn with a $B \times T$ template, based on 300 training utterances per speaker. The procedure is described in earlier work [3, 6]. The concatenated 8500 $(34 \cdot 250)$ atom speech basis was used to factorise the remaining 200 training utterances for learning the activation-state mapping matrices needed for sparse classification, in each case with factorisation parameters matching the corresponding test set-up. Mappings were learnt with ordinary least squares regression. During development and test set recognition, a 250-atom noise basis was sampled for each utterance from its noise context and added to the total basis [3].

The binary group sparsity matrix $\mathbf{G}_{\mathrm{B}}$ ($S \times L$, $S = 34$, $L = 8500 - 8750$) was simply set to 1 for atoms corresponding to speaker s, in other words, for entries 1–250 of group (row) 1, entries 251–500 of group 2 and so forth. The noise atoms at indices 8501–8750, used in all noisy test conditions, did not belong to any group, i.e., the group sparsity constraint was not used for noise. L_1 sparsity weights in matrix $\mathbf{\Lambda}_1$ were kept at 0.1 for entries corresponding to speech and 0.85 for noise as in earlier work. Group sparsity weight λ_{g} was set to 0.1 based on development set factorisation. All sparsity weights were multiplied by the mean of 1-norms of dictionary atoms to tie the relative weights of KL-divergence and sparsity costs together.

4.3. Recognition experiments

The 3600 test utterances were factorised using the joint 8750-atom basis and 300 iterations of the update rule given in Equation (9). Activation matrices were used for three evaluations:

1. Speaker identification
2. Speech recognition in an external GMM back-end via feature enhancement
3. Speech recognition by sparse classification, that is, determining the state likelihoods from activation weights

All experiments were run with and without the group sparsity penalty, all other parameters remaining identical.

Speaker identification was performed using sparse discriminant analysis (SDA) [6, 13]. Considering the fact that there is only one speaker present in an utterance, we used the summed activity vector $\mathbf{x}_{\Sigma}$ over an utterance as a feature vector. In order to make the vector invariant to different utterance lengths, the vectors were normalised by the number of windows. The feature vectors from 200 training files per speaker were supplied to an SDA algorithm to find the sparse directions with maximum separability between speakers and minimum variability within speakers. By projecting the 200 vectors from each speaker on sparse discriminant directions, an average model of a speaker was made by simply averaging them. The activity vector of a test segment was also mapped onto SDA directions and dot scoring was employed as the speaker identification score. The number of non-zero elements in SDA was set to 500.

For GMM-based speech recognition, we used the CHiME HTK language model, multi-condition trained GMMs [2], and feature enhancement as in previous work [3]. True speaker identity was exploited in GMM selection in the back-end.

Sparse classification was also performed as in earlier work [3]. Speaker-dependent models were used for Viterbi decoding, although their contribution is limited to transition probabilities, which are highly similar for all speakers.

Table 1: Speaker identification rate (%) comparison for no group sparsity constraint ($\lambda_{\mathrm{g}} = 0$) and with group sparsity ($\lambda_{\mathrm{g}} = 0.1$) on the CHiME test set.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
$\lambda_{\mathrm{g}} = 0$	99.8	99.3	98.7	95.5	94.3	82.5	95.0
$\lambda_{\mathrm{g}} = 0.1$	99.7	99.3	99.2	96.7	93.7	85.7	95.7

5. Results

Results for speaker identification are shown in Table 1. Rates without using group sparsity are shown on the first line, and rates with group sparsity enabled on the second. We observe 0.7% absolute (14% relative) improvement in the average score. Individual SNR scores vary to both directions with debatable significance considering the 600 utterance set size. More on factorisation-based speaker identification results including comparison with GMM baseline can be found in [6].

Table 2 shows the results of speech recognition using factorisation-based enhancement and a GMM back-end. The first two rows contain unenhanced baseline scores for the clean-trained CHiME standard models [12] and multi-condition (MC) trained models [2]. Results for enhancement with different factorisation models are given in the second part of the table. The 8500-atom multi-speaker basis is employed first with L_1 sparsity only, and then with group sparsity enabled. To evaluate the 'oracle' performance obtainable by perfect speaker discrimination, the results on the last row use the true speaker's 250-atom speech basis and the same 250 noise atoms to enhance the signals. We notice that adding group sparsity to multi-speaker basis enhancement produces slight improvements, but only in the noisy end and by a small margin. Neither variant manages to match oracle single-speaker enhancement.

Sparse classification results can be found in Table 3. The same factorisation variants as in enhancement are used for evaluation. This time group sparsity improves the multi-speaker factorisation scores significantly, making them comparable to oracle single-speaker factorisation and classification.

6. Discussion

The results for speaker identification (Table 1) are not entirely conclusive. However, the -6 dB condition is of special interest, because many of its utterances contain loud non-target speech as their background noise. The 18% relative improvement there suggests, that sharpening the distribution of speaker activity manages to remove some interference from non-target speakers. Clean end results are near-perfect to begin with, and there is little confusion between speakers. Consequently no significant changes take place there. Due to the novelty of the approach, further test should be conducted for more conclusive results.

In factorisation-based speech enhancement (Table 2), the speaker identity and state information of atoms is not used in any way — only the spectral features. Therefore features from another speaker are equally valid as long as the spectrograms match, and group sparsity has a limited effect. Improvements in the noisy end can probably be attributed to the non-target background speakers, and the restricted dictionaries' ability to reject secondary identities matching to them. Due to stronger discrimination, such speech is more likely to become modelled with noise atoms as expected. Again, in the clean end differences are limited to only a few test files.

In sparse classification (Table 3), state likelihoods are acquired solely from activation weights and atom labelling. Be-

Table 2: Enhancement-based speech recognition scores (%) over SNRs. Results are shown for clean-trained CHiME baseline models, multi-condition (MC) trained models without enhancement, multi-speaker (MS) enhancement either without or with group sparsity, and finally enhancement by only using the true, single speaker's basis (SS).

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
GMM baseline scores without enhancement							
CHiME	82.4	75.0	62.9	49.5	35.4	30.3	55.9
MC	91.3	86.8	81.7	72.8	61.1	54.5	74.7
GMM recognition with MC models and enhancement							
MS, $\lambda_{\mathrm{g}} = 0$	92.6	90.3	88.2	84.5	75.6	69.8	83.5
MS, $\lambda_{\mathrm{g}} = 0.1$	92.4	90.4	88.0	85.3	76.2	70.4	83.8
SS	93.0	91.2	90.0	85.2	79.0	72.9	85.2

cause speaker models are trained independently, activations of atoms from other speakers introduce unreliable factors to the final likelihoods. Group sparsity reduces such errors by favouring small sets of active speakers. It is noteworthy that our multi-speaker basis with group sparsity produces recognition rates closely matching informed recognition using the true speaker's basis alone. Because the HMMs can be trained speaker-independently, the whole recognition process becomes speaker-independent over the set of modelled speakers. Together with a robust speaker identification algorithm, the framework provides reliable classification results for both speaker identity and the phonetic content in a scenario, where one unknown speaker from multiple candidates is active at a time.

Concerning the overall rates, it should be noted that the presented framework used small 250-atom speech and noise bases. In other work, we have presented several alternatives for speech and noise modelling [3]. Better results could be achieved by using more accurate speech and noise models, although the efficiency of improved models in conjunction with group sparsity needs to be investigated. While in the presented results speech enhancement was found to perform better than sparse classification, for different bases and features the order may become reversed [3]. Moreover, the two approaches have been found to complement each other in multi-stream recognition [14].

In this study, group sparsity was used for speaker discrimination. However, it is equally feasible to select any sets of atoms for the groups based on their expected co-occurrence. The atom weights in groups need not to be binary either. Different temporal spans can be selected for groups either by choosing an appropriate factorisation spectrogram length, or adjusting the window span used in Equation (7), and then spreading the group sparsity penalty vector (11) accordingly.

7. Conclusions

We proposed using group sparsity in addition to L_1 sparsity in spectral factorisation based noise robust speech recognition in order to limit the number of active speakers from multiple candidates. An iterative update rule was presented for solving convolutive non-negative matrix factorisation with consistent group sparsity over all time indices in an utterance. We found out that the new model manages to narrow down the distribution of speakers, producing marginal but consistent improvements in speaker and speech recognition results. The presented model is generic and allows enforcing also other kinds of group structures in dictionary-based audio spectrogram factorisation.

Table 3: Speech recognition scores (%) with sparse classification. Results are shown for the multi-speaker (MS) basis without and with group sparsity, and then for using the true, single speaker only (SS).

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Sparse classification scores							
MS, $\lambda_{\mathrm{g}} = 0$	89.3	87.7	81.5	78.0	68.1	57.9	77.1
MS, $\lambda_{\mathrm{g}} = 0.1$	90.4	88.4	85.7	80.8	73.4	64.3	80.5
SS	89.8	89.0	84.3	81.8	73.9	65.8	80.8

8. Acknowledgements

The work of Rahim Saeidi was funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 238803.

9. References

[1] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[2] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 24–29.

[3] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition," *submitted work*, 2012.

[4] S. Rennie, J. Hershey, and P. Olsen, "Single Channel Multi-talker Speech Recognition: Graphical Modeling Approaches," *IEEE Signal Processing Magazine, Special Issue on Graphical Models*, vol. 27, 2010.

[5] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[6] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen, "Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification," in *Odyssey speaker and language recognition workshop*, Singapore, 2012.

[7] S. Bengio, F. C. N. Pereira, Y. Singer, and D. Strelow, "Group Sparse Coding," in *Proc. NIPS*, 2009, pp. 82–89.

[8] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito Nonnegative Matrix Factorization with Group Sparsity," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 21–24.

[9] L. Meier, S. Van De Geer, and P. Bühlmann, "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society: Series B*, vol. 70, pp. 53–71, 2008.

[10] Q. Tan and S. Narayanan, "Novel Variations of Group Sparse Regularization Techniques with Applications to Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1337–1346, 2012.

[11] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Non-negative Matrix Deconvolution in Noise Robust Speech Recognition," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4588–4591.

[12] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME Corpus: a Resource and a Challenge for Computational Hearing in Multisource Environments," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1918–1921.

[13] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse Discriminant Analysis," *Technometrics*, vol. 54, no. 4, pp. 406–413, 2011.

[14] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative Matrix Factorization for Highly Noise-robust ASR: To Enhance or to Recognize?," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 4681–4684.