Chapter 6

Dictionary Learning-Based Speech Enhancement

Viet-Hang Duong, Manh-Quan Bui and
Jia-Ching Wang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.85308

Abstract

This chapter presents an overview of dictionary learning-based speech enhancement methods. Specifically, we review the existing algorithms that employ sparse representation (SR), nonnegative matrix factorization (NMF), and their variations applying for speech enhancement. We emphasize that there are two stages in a speech enhancement system, namely learning dictionary and enhancement. The two scenarios of learning dictionary process, offline and online, are discussed carefully as well. We finally present some evaluation methods and suggest the future lines of work.

Keywords: dictionary learning, nonnegative matrix factorization, projected gradient descent, speech enhancement, sparse representation

1. Introduction

Speech is the most important tool of expression and it is crucial information carrier of language communication. Speech signals in real-world scenarios are corrupted due to some disturbing noise such as background noise, reverberation, babble noise, etc. The purpose of speech enhancement (SE) is to extract the clean speech signal from the interferer components mixture as much as possible, so as the clarity and intelligibility of the speech signal. The research of speech enhancement technology is particularly important and difficult. Speech denoising is an importance problem with increasing various applications as hearing aids, speech/speaker recognition, mobile communications over telephone, and Internet [1]. The difficulties arise from the nature of real-world noise that is often unknown, nonstationary, potentially speech-like, overlapping between [1–3].

Assume that the noisy speech x is a linear additive mixture of the clean speech s and the interfere n as defined in the following equation:

$$x(t) = s(t) + n(t) \tag{1}$$

where $x(t)$ is the time-domain mixture signal at sample t, and $s(t)$ and $n(t)$ are the time-domain speech and interferer signals, respectively. The speech enhancement algorithm attempts to suppress noise without distorting speech and obtain the enhanced speech components $\hat{s}$ from the noisy signal and reconstruct the original clean speech. In other words, speech enhancement algorithms try to reduce the impact of background noise on the speech signal. Most traditional speech enhancers are implemented in the short-time Fourier transform (STFT) domain with $\mathbf{X} = |STFT\{x(t)\}|^{\gamma}$ where γ = 1 gives the magnitude of spectrum or the power spectrum by γ = 2. The inverse Fourier transformation then is used to convert the estimated speech to the time domain, assuming that the phase of the interferer can be approximated with the phase of the mixture [4].

The speech enhancement techniques mainly focus on removal of noise from speech signal. The various types of noise and techniques for removal of those noises are presented [5–13]. The famous spectral subtraction technique [5] extracted the clean speech spectrum based on the principle that the noise contamination process is additive. The major advantage of the spectral subtraction method is their simplicity by subtracting an estimation of the interfere spectrum from the observed mixture spectrum [5, 6]. The main problem with the magnitude spectral subtraction is that it does not attenuate noise sufficiently negative magnitude by error in the subtraction.

Filtering techniques [7, 8] or short-time spectral amplitude (STSA) estimators [9] or estimators based on super-Gaussian prior distributions for speech DFT coefficients are [10–13] the statistical models assumed for each of the speech and noise signals that estimate the clean speech from the noisy observation without any prior information on the noisy type or speaker identity. However, in the case of nonstation of background noise, these methods face much difficulty in estimating the noise power spectral density (PSD) [14–16].

Recently, dictionary learning (DL) techniques, which build dictionary consisting of atoms and represent a class of signals in terms of the atoms, have been shown to be effective in machine learning, neuroscience, and audio processing [17–20]. In speech enhancement, the dictionary models utilize specific types of the a priori information considered for both the speech and noise signals [21–25]. This class of methods assumes that a target spectrogram can be generated from a set of basis target spectra (a dictionary) through weighted linear combinations. Generally, this approach decomposes the time-frequency representations (the power or magnitude spectrogram) of noisy speech in terms of elementary atoms of a dictionary. One of the key issues in dictionary-based speech enhancement is how to precisely learn a dictionary. Dictionary learning methods are commonly based on an alternating optimization strategy, in which the signal representation is fixed, and the dictionary elements are learned; then the sparse signal representation is found, while the dictionary is fixed. Two popular methods have appeared to determine a dictionary within a matrix decomposition including sparse coding [26] and nonnegative matrix factorization (NMF) [27].

The observation that speech and other structured signals can be well approximated by few atoms of a suitably trained dictionary [28], which lies at the core of sparse representation (SR). In SR, sparse signals can be reconstructed with a few atoms of an overcomplete dictionary. Recently, developed SR has been shown to be effective in data representation, which factorizes given matrix with regularization methods or regularization term to constrain the sparsity of desire representation. Since speech signals are generally sparse in the time-frequency domain and many types of noise are nonsparse, the target speech signal was decomposed and reconstructed from the noisy speech-driven sparse dictionary [21–23].

In many reality applications, the nonnegativities of the signals and the dictionary are required such as multispectral data analysis [29, 30], image representation [31, 32], and some other important problems [33, 34], the so-called nonnegative dictionary learning becomes necessary. Nonnegative matrix factorization is a popular dictionary method, which projects the given nonnegative matrix onto the subspace spanned by nonnegative dictionary vectors. Treating speech enhancement as a source separation problem between speech and noise, NMF-based techniques can be used to factorize spectrograms into nonnegative speech and noise dictionaries and their nonnegative activations. On the one hand, a clean speech signal can be estimated from the product of speech dictionaries and their activation.

In this chapter, we review the dictionary learning approaches for speech enhancement. After a brief introduction to the problem and its characterization as a sound source separation task, we present a survey on both theoretically and applicable of dictionary-based techniques, the main subject of this chapter. We finally provide an overview of the evaluation methods and suggest some future lines of works.

2. Background

Dictionary learning performs approximate matrix factorization of a data matrix into the product of a dictionary matrix and a coding matrix, under some sparsity constraints on the coding matrix. Dictionary learning is the generalization of gain-shape codebook learning. Signal vectors are represented as linear combinations of multiple dictionary atoms, allowing for lower approximation error while maintaining equal dictionary size. Two relatively different methods are described for how to form the dictionary from the given data including sparse representation (SR) and nonnegative matrix factorization (NMF).

2.1. Sparse representation (SR) and K-SVD algorithm

Let $\mathbf{X}$ be a matrix of M training signals $\mathbf{X} = \{\mathbf{x}_m\}_{m=1}^{M} \in \mathbb{R}^N$. SR dictionary learning framework consists in finding a dictionary $\mathbf{D}$ of K unit-norm atoms $\mathbf{D} = \left[d_{(1)} \ldots d_{(K)}\right] \in \mathbb{R}^{N \times K}$ and sparse coefficients $\mathbf{C} = \{\mathbf{c}_m\}_{m=1}^{M} \in \mathbb{R}^K$ such that the approximation error between $\mathbf{X}$ and $\mathbf{DC}$ is sufficiently small. For example, if the exact sparsity level T_0 is known, the problem can be formalized as minimizing the error cost function $O_{SR}(\mathbf{D}, \mathbf{C})$ defined as:

$$f_{SR}(\mathbf{D},\mathbf{C}) = \|\mathbf{X} - \mathbf{DC}\|_F^2, \quad \text{s.t.} \forall i, \|\mathbf{c}_i\|_0 \le T_0 \tag{2}$$

where $\|.\|_F$, $\|.\|_0$ denote the Frobenius and l_0 norm, respectively.

Eq. (2) shows that a signal **x** can be expressed as the linear combination of only a few column vectors in **D**. Matrix factorization problem (2) is a difficult problem, since the joint optimization of **D** and **C** is nonconvex. Many dictionary algorithms follow an iterative scheme that alternates between updates of dictionary **D** and sparse coding **C** to minimize the cost function (2). K-SVD, one of the methods, goes under the category of sparse representation (SR), which came from the theory of sparse and redundant representation of signals. It was first introduced by Aharon et al. [34]. The K-SVD algorithm defines an initial overcomplete dictionary matrix $\mathbf{D}_0 \in \mathbb{R}^{N\times K}$ and operates alternating two step iterations between optimizing the coding and the dictionary as follows:

The sparse coding approximation step derives the column $\mathbf{c}_m$, $m = 1. M$ by using the orthogonal matching pursuit (OMP) algorithm with given **X** and **D** to solve the following equation:

$$\text{argmin}\|\mathbf{c}_m\|_0 \quad \text{s.t} \quad \|\mathbf{x}_m - \mathbf{Dc}_m\|_2 \le \sigma \tag{3}$$

The updating dictionary step is taken by minimizing the approximation error (2) with the current coding **C**. Atom-by-atom is updated in an iterative process.

$$\text{Because } \|\mathbf{X} - \mathbf{DC}\|_F^2 = \left\|\mathbf{X} - \sum_{i=1}^{K} \mathbf{d}_i\mathbf{c}^{[i]}\right\|_F^2 = \left\|\left(\mathbf{X} - \sum_{i\neq j} \mathbf{d}_i\mathbf{c}^{[i]}\right) - \mathbf{d}_j\mathbf{c}^{[j]}\right\|_F^2 = \left\|\mathbf{R}^{(j)} - \mathbf{d}_j\mathbf{c}^{[j]}\right\|_F^2 \tag{4}$$

where $\mathbf{c}^{[i]}$ is the ith row of **C**. The residual norm is minimized by seeking for a rank-one approximation [35]. The approximation is based on computing the singular value decomposition (SVD) [23].

2.2. Nonnegative matrix factorization (NMF) theory

Nonnegative matrix factorization (NMF) can be viewed as an approach for dictionary learning. NMF, first introduced by Paatero and Tapper [36] and later popularized by Lee and Seung [23, 27–37], has been known as a part-based representation model. Different to other matrix factorization approaches, NMF takes into account the fact that most types of real-world data, particularly sound and videos, are nonnegative and maintain such nonnegativity constraints in factorization. Moreover, the nonnegativity constraints in NMF are compatible with the intuitive notion of combining parts to form a whole, that is, they provide a parts-based local representation of the data. A parts based model not only provides an efficient representation of the data but can potentially aid in the discovery of causal structure within it and in learning relationships between the parts.

Given a nonnegative matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] \in \mathbb{R}_+^{N\times M}$, a positive integer $K << \min\{N, M\}$, NMF projects **X** onto a space by a linear combination of a set of nonnegative basis vectors $\mathbf{D} = \{\mathbf{d}_{nk}\}$, that is, $\mathbf{X} \approx \mathbf{DC}$ where $\mathbf{C} = \{\mathbf{c}_{km}\}$, $\mathbf{c}_{km} \ge 0$. In order to find an approximate factorization

for the matrix **X**, cost function that quantifies the quality of the decomposition needs to be defined. Operationally, NMF can be described as the following objective function

$$\min_{\mathbf{D},\,\mathbf{C}\geq 0} f(\mathbf{X}\|\mathbf{DC}) \tag{5}$$

where f is denoted a distance metric.

Different the similarity measures between **X** and the product **DC** lead to different variants of NMF. The common choices include Euclidean distance [38], generalized Kullback-Leibler divergence [39], Itakura-Saito divergence [40]… For instance, the NMF based on Kullback–Leibler (KL) divergence is formulated as follows:

$$f_{KL}(\mathbf{X},\mathbf{DC}) = \sum_{i,j}\left(\mathbf{x}_{ij}\log\frac{\mathbf{x}_{ij}}{(\mathbf{DC})_{ij}} - \mathbf{x}_{ij} + (\mathbf{DC})_{ij}\right) \tag{6}$$

There exist different optimization models for the approximation factorization (5) [36, 39, 40]. The most popular solution is alternative multiplicative update rules (MURs) [36], which do not have required user-specified optimization parameters. For a KL cost function (6), the iteratively updating rules are given by:

$$\mathbf{c}_{a\mu} \leftarrow \mathbf{c}_{a\mu}\frac{\sum_i \mathbf{d}_{ia}\mathbf{x}_{i\mu}/(\mathbf{DC})_{i\mu}}{\sum_t \mathbf{d}_{ta}} \tag{7}$$

$$\mathbf{d}_{ia} \leftarrow \mathbf{d}_{ia}\frac{\sum_\mu \mathbf{c}_{a\mu}\mathbf{X}_{i\mu}(\mathbf{DC})_{i\mu}}{\sum_s \mathbf{c}_{as}}; \tag{8}$$

However, it is found that the monotonicity guaranteed by the proof of multiplicative updates may not imply the full Karush-Kuhn-Tucker conditions [39, 40]. MUR is relatively simple and easy to implement, but it converges slower in comparison with gradient approaches [41]. More efficient algorithms equipped with stronger theoretical convergence property have been introduced. One popular method is to apply gradient descent algorithms with additive update rules, which are represented by the projective gradient descent method (PGD) [42]. In PGD framework, to select the learning step size, a line search method with the Armijo rule is applied [42] and the new estimate is obtained by first calculating the unconstrained steepest-descent update and then zeroing its negative elements. In addition, considering the separate convexity, the two-variable optimization problem is converted into the nonnegative least squares (NLS) optimization subproblems, which alternate the minimization over either **D** or **C**, with the other matrix fixed.

Because of the initial condition $K << \min\{N, M\}$, the obtained basis vectors are incomplete over the original vector space. In other words, this NMF approach tries to represent the high-dimensional stochastic pattern with far fewer bases, so the perfect approximation can be achieved successfully only if the intrinsic features are identified in **D**.

NMF will not get the unique solution under the sole nonnegativity constraint. Hence, to remedy the ill-posedness, it is imperative to introduce additional auxiliary constraints on **D**

and/or **C** as regularization terms, which will also incorporate prior knowledge and reflect the characteristics of the issues more comprehensively. The constrained NMF models can be unified under the similar extended objective function

$$\min_{\mathbf{D},\,\mathbf{C}\geq 0} f_{constrainedNMF}\left(\mathbf{X}\middle\|\mathbf{DC}\right) = \min_{\mathbf{D},\,\mathbf{C}\geq 0}\left[f(\mathbf{X}\|\mathbf{DC}) + \alpha g(\mathbf{D}) + \chi h(\mathbf{C})\right] \tag{9}$$

where the regularization parameters $\alpha$and χ are used to balance the trade-off between the fitting goodness and the constraints $g(\mathbf{D})$ and $h(\mathbf{C})$.

The performance of NMF can be improved by imposing extra constraints and regularizations. For the sparseness learning, the sparse term $h(\mathbf{C})$ expects to constraint the mount of nonzero elements in each column of the projection matrix. The L_0 norm could be selected to count nonzero elements in **C** [43]. One limitation of using L_0 norm is that the solution is not unique because of many local minima of the cost function. In this situation, the L_1 norm of the projection matrix is usually replaced as a relaxation of the L_0 penalty [44, 45].

$$\|\mathbf{C}\|_1 = \sum_{j=1}^{M}\left\|\mathbf{c}_{:j}\right\|_1 = \sum_{j=1}^{M}\left(\sum_{i=1}^{K}\left|\mathbf{c}_{ij}\right|\right) \tag{10}$$

3. Dictionary learning-based speech enhancement

A major outcome of speech enhancement techniques is the improved quality and reduced listening effort in the presence of an interfering noise signal. The decomposition of time-frequency representations, such as the power or magnitude spectrogram in terms of elementary atoms, has become a popular tool in speech enhancement since their success in finding high-"quality" dictionary atoms that best describe latent features of the underprocessed data. The dictionary-based techniques utilize specific types of the a priori information of speech or noise [21, 23, 46–50]. A priori information can be typical patterns or statistics obtained from a speech or noise database. Dictionary-based speech enhancement consists of two separate stages: a training stage, in which the model parameters are learned, and a denoising stage, in which the noise reduction task is carried out. In the first step, dictionary **D** is learned while fixing coefficient matrix **C**, and in second step, **C** is computed with the fixed dictionary matrix **D**. This process of alternate minimization is repeated iteratively until a stopping criterion is reached. In order to learn dictionary atoms capable of revealing the hidden structure in speech, long temporal context of speech signals must be considered. Two major classes of dictionary-based speech enhancement techniques may be the offline learning and online learning. Offline algorithms for dictionary learning are second-order iterative batch procedures, accessing the whole training set at each iteration in order to minimize a cost function under some constraints [21–23]. In speech enhancement, learning spectrotemporal atoms spanning several consecutive frames is done through training large volumes of datasets, which places unrealistic demand on computing power and memory. In large-scale tasks, online dictionary learning tends to gain lower empirical cost than conventional batch learning [46–50].

Speech enhancement herein is implemented in the short-time Fourier transform (STFT) magnitude domain, assuming that the phase of the interferer can be approximated with the phase of the mixture. The number of frequency bins per frame is determined by the length of the time-domain analysis window, where a Hamming window was chosen for the STFT. The temporal smoothness frames are determined by the time-domain analysis window overlap, where a minimum amount of overlap is necessary to avoid aliasing.

3.1. Offline dictionary

Sparse representation has been described as an overcomplete models wherein the number of bases is greater than the dimensionality of spectral representations. In sparse representation, sparse signals can be expressed as the linear combination of only a few atoms in an overcomplete dictionary. While speech signals are generally sparse in the time-frequency domain and many types of noise are nonsparse, the target speech signal reconstructed from the noisy speech is considered as clean speech. A possibly overcomplete dictionary of atoms is trained for both speech and interferer magnitudes, which are then concatenated into a composite dictionary. The training process of updated dictionary is drawn in **Figure 1**.

When applying the sparse coding technique to speech enhancement, it is desirable to have the trained offline clean speech dictionary $\mathbf{D}_{speech}$ to be coherent to the speech signal and incoherent to the background noise signal as well as a coherent noise dictionary $\mathbf{D}_{noise}$. In the enhancement step, the noisy speech is sparsely coded in the composite dictionary $[\mathbf{D}_{speech}, \mathbf{D}_{noise}]$. As a result, this mixture of speech and interferer $\mathbf{x}$ is explained by a sum of a linear combination of atoms from the speech dictionary $\mathbf{D}_{speech}$ and a linear combination of atoms from the interferer dictionary $\mathbf{D}_{noise}$. The noisy $\mathbf{x}$ is coded using the least angle regression (LASSO) [51] with a preset threshold θ as follows:

$$\arg \min_{\mathbf{c}^{speech}, \mathbf{c}^{noise}} \left\| \mathbf{x} - \left[\mathbf{D}_{speech}\ \mathbf{D}_{noise}\right] \begin{bmatrix} \mathbf{c}_{speech} \\ \mathbf{c}_{noise} \end{bmatrix} \right\|_2 s.t. \frac{\|\mathbf{c}\|_1}{\|\mathbf{x}\|_2} \le \theta \tag{11}$$

The clean speech magnitude is estimated by disregarding the contribution from the interferer dictionary, preserving only the linear combination of speech dictionary atoms (analogously for the interferer) and

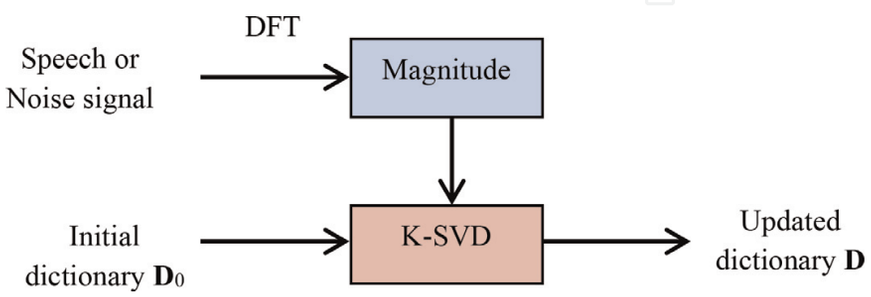

Figure 1. The training process of updated dictionary.

$$\hat{s} = \mathbf{D}_{speech}\mathbf{c}_{speech} \tag{12}$$

It is known that NMF represents data as a linear combination of a set of basis vectors, in which both the combination coefficients and the basis vectors are nonnegative. Although the basis learned by NMF is sparse, it is different from sparse coding [26]. This is because NMF learns a low rank representation of the data, while sparse coding usually learns the full rank representation. Treating speech enhancement as a source separation problem (speech and noise), NMF-based techniques can be used to factorize spectrograms into nonnegative speech and noise dictionaries and their nonnegative activations. Assume that a clean speech spectrogram as $\mathbf{X}_{speech}$ and a clean noise spectrogram as $\mathbf{X}_{noise}$. Consider a supervised denoising approach where the clean speech basis matrix $\mathbf{D}_{speech}$ and the clean noise basis matrix $\mathbf{D}_{noise}$ are learned separately by performing NMF on the speech and the noise. During training process, minimized $f\left(\mathbf{X}_{speech}\middle\|\mathbf{D}_{speech}\mathbf{C}_{speech}\right)$ and $f\left(\mathbf{X}_{noise}\|\mathbf{D}_{noise}\mathbf{C}_{noise}\right)$ are employed.

To reduce the noise in the noisy speech, the concatenated dictionary $\mathbf{D} = [\mathbf{D}_{speech}, \mathbf{D}_{noise}]$ is fixed and utilized in decomposing the noisy speech $\mathbf{X}_{noisy}$ by

$$\min_{\mathbf{C}_{noisy} \geq 0} f\left(\mathbf{X}_{noisy}\middle\|\mathbf{D}\mathbf{C}_{noisy}\right) \tag{13}$$

where the time-varying activation matrix is formulated $\mathbf{C}_{noisy} = \begin{bmatrix} \mathbf{C}'_{noise} \\ \mathbf{C}'_{speech} \end{bmatrix}$.

Discarding the noise coding matrix, the target speech is estimated from the product of speech dictionaries and their activations as

$$\widehat{\mathbf{X}}_{speech} = \mathbf{D}_{speech}\mathbf{C}'_{speech} \tag{14}$$

The clean speech waveform is estimated using the noisy phase and inverse DFT and the general framework of NMF-based speech enhancement is drawn in **Figure 2**.

3.2. Online dictionary learning

The aforementioned dictionary learning approaches access the whole training set to determine the bases, which are referred as offline training process. These methods were reported to have good performance on modeling nonstationary noise types, which had been seen during training. For the time-frequency analysis of audio signals, however, the obtained basis may not be adequate to capture the temporal dependency of repeating patterns within the signal, and the success of these methods strongly relies on the prior knowledge of noise or speech or both, which limits implementations of the models. Recently, the online dictionary learning methods have been proposed in two aspects of implementing scheme [46–50] and circumventing the mismatch problem between the training and testing stages [24, 52].

One drawback of the multiplicative update procedure on offline dictionary learning is the requirement of all the training signals to be read into memory and processed in each iteration.

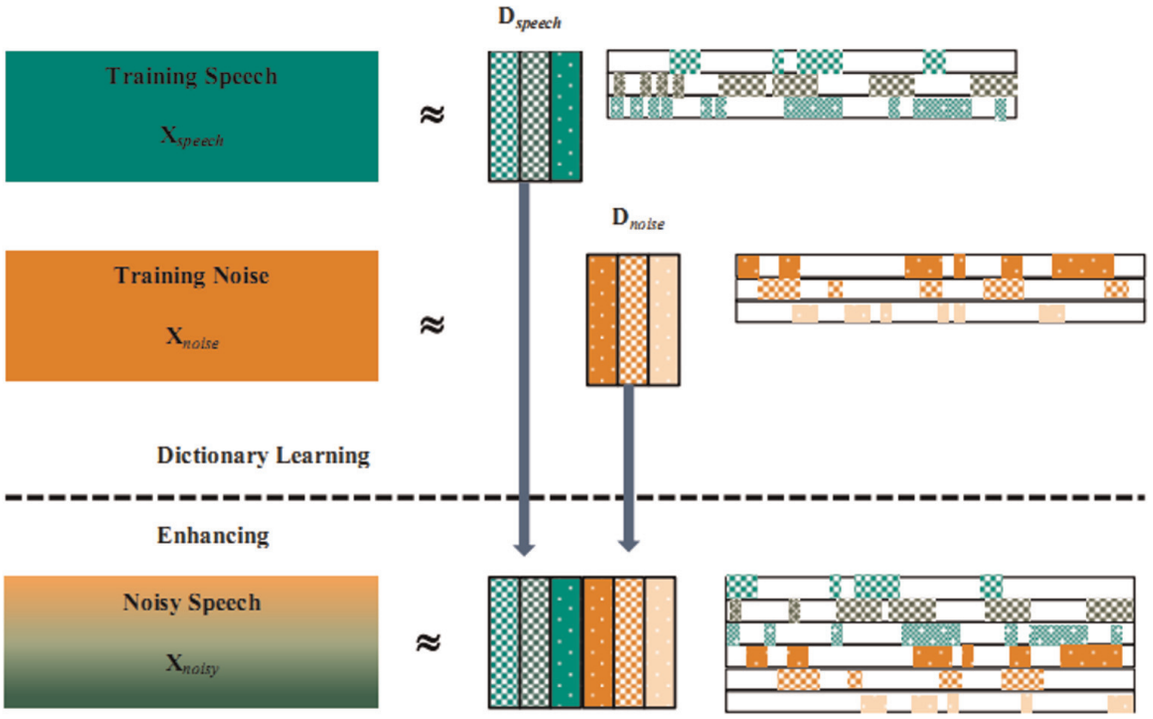

Figure 2. Block diagram of NMF-based speech enhancement.

This high demand on both computing resources and memory is prohibitive in large-scale tasks. To address this problem, the online optimization algorithms were developed in an incremental fashion, which processes one sample of the training set at a time based on stochastic approximations or only a part of the training data at a time and updates patterns gradually until completely processed whole training corpus [46–48, 51]. More specifically, given M samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\} \in \mathbb{R}_+^N$ distributed in the probabilistic space $\wp \in \mathbb{R}_+^N$, the conventional NMF learns subspace $Q \subset \wp$ spanned by a base $\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K\} \in \mathbb{R}_+^N$ and satisfies the expected cost:

$$\min_{\mathbf{D} \in \mathbb{R}_+^{N \times K}} \sum_{i=1}^{M} f(\mathbf{x}_i \| \mathbf{D}\mathbf{c}_i) \text{ with fixed } \mathbf{c}_i \tag{15}$$

$$\text{or} \quad \min_{\mathbf{D} \in \mathbb{R}_+^{N \times K}} E_{\mathbf{x}_i \in \wp}(f(\mathbf{x}_i \| \mathbf{D}\mathbf{c}_i)) \tag{16}$$

where $E_{\mathbf{x}_i \in \wp}$ denoted the expectation on $\wp$.

The coefficient matrix is computed by

$$\min_{\mathbf{C} \in \mathbb{R}_+^{K \times M}} f(\mathbf{X} \| \mathbf{D}\mathbf{C}) \tag{17}$$

For the online NMF framework, at step t, on the arrival of sample $\mathbf{x}^{(t)}$, the corresponding coefficient $\mathbf{c}^{(t)}$ is formulated by

$$\min_{\mathbf{c}^{(t)} \in \mathbb{R}_+^K} f\left(\mathbf{x}^{(t)} \middle\| \mathbf{D}^{(t-1)}\mathbf{c}^{(t)}\right) \tag{18}$$

where $\mathbf{D}^{(t-1)}$ is the previous basis matrix. The matrix $\mathbf{D}^{(t)}$ is updated by

$$\mathbf{D}^{(t)} = \arg \min_{\mathbf{D} \in \mathbb{R}_+^{N \times K}} E_{\mathbf{x} \in \wp^{(t)}} \left(f(\mathbf{x} \| \mathbf{D}\mathbf{c})\right) \tag{19}$$

where $\wp^{(t)} \subset \wp$ is the probabilistic subspace spanned by the arrived elements $\left\{\mathbf{x}^{(1)}, , \mathbf{x}^{(2)}, \ldots, , \mathbf{x}^{(t)}\right\} \in \mathbb{R}_+^N$ and the corresponding $\left\{\mathbf{c}^{(1)}, , \mathbf{c}^{(2)}, \ldots, , \mathbf{c}^{(t)}\right\} \in \mathbb{R}_+^K$ are computed available in the previous t steps.

In [50], an online noise basis learning scheme is proposed that uses the temporal dependencies of speech and noise signal to construct informative prior distribution. In this model, the noise basis matrix is learned from the noisy observation. To update the noise basis, the past noisy DFT magnitude frames are stored into a buffer and the buffer will be then updated with fixed speech basis when a new noisy frame arrives.

Kwon et al. [52] present a speech enhancement technique combining statistical models and NMF with online update of speech and noise bases. A cascaded structure of combining a statistical model-based enhancement (SE) (the first state) [53] and NMF approach (second stage) with simultaneous update of speech and noise bases is proposed. In this model, the output clean speech at current frame is fed as an input to update the speech and noise bases in the following frame. In other words, at each frame, the clean speech estimation is obtained; the speech and noise bases for the NMF analysis in the following frame are updated. This online bases update makes it possible to deal with the speech and noise variations that cannot be covered by the training noise database and is considered a promising way to cope with the nonstationary nature of the signal. The noisy data $\mathbf{X}'(t)$ used for the online bases update herein is constructed by concatenating preenhanced output $\mathbf{X}_{SE}(t)$ of performing statistical model-based enhancement (SE) with the current frame input $\mathbf{X}(t)$. The updating dictionary process will be learned by adding a regular term to the original objective function as follows:

$$f_{onlineSE+NMF}\left(\mathbf{X}'(t) \middle\| \mathbf{D}'(t)\mathbf{C}'(t)\right) = f\left(\mathbf{X}'(t) \middle\| \mathbf{D}'(t)\mathbf{C}'(t)\right) + \alpha \|\mathbf{D}(t) - \mathbf{D}'(t)\|^2 \tag{20}$$

where $\mathbf{D}'(t) = [\mathbf{D}'_{speech}, (t)\mathbf{D}'_{noise}(t)]$ denotes the basis matrix in NMF decomposing of the concatenated noisy data $\mathbf{X}'(t)$ and $\mathbf{D}(t) = [\mathbf{D}_{speech}, (t)\mathbf{D}_{noise}(t)]$ is the basis matrix used to analyze the t-frame $\mathbf{X}(t)$ in the second state.

4. Summary and discussion

In the experimental simulations, speech and noise materials were selected from TIMIT [53] (192 sentences), NOISEX-92 DBs (15 types of noise: birds, casino, cicadas, computer keyboard, eating chips, f16, factory1, factory2, frogs, jungle, machineguns, motorcycles, ocean, pink, and volvo) [54], the GRID audiovisual corpus (34 speakers of both genders) [55], the NOIZEUS

speech corpus (30 utterances with clean samples) [1]. The noisy speech examples were synthesized by adding clean speech to different types of noises at various input SNRs.

Speech enhancement algorithms aim to improve both the speech quality and the speech intelligibility. A high-quality speech signal is perceived as being natural and pleasant to listen to, and free of distracting artifacts. An effective technique should suppress noises without bringing too much distortion to the enhanced speech. Measuring speech quality is challenging, as it is subjective and can be classified into subjective and objective measures. The speech enhancement performance was commonly evaluated in terms of three criteria including the signal to noise ratio (SNR) of enhanced speech [56], the segmental SNR (segSNR) [56], or the perceptual estimation of speech quality score (PESQ) [57–59]. Given the true and estimated speech magnitude spectra, the frequency-weighted segmental SNR is defined as:

$$SNR = 10 \times \log \left(\frac{\sum_t \left(\mathbf{X}_{noisy}(t) - \mathbf{X}_{speech}(t) \right)^2}{\sum_t \left(\widehat{\mathbf{X}}_{speech}(t) - \mathbf{X}_{speech}(t) \right)^2} \right) \tag{21}$$

segSNR is a conceptually simple objective measure, computed on individual signal frames, and the per-frame scores are averaged over time.

$$segSNR = \frac{1}{N} \sum_{b=1}^{N} 10 \times \log \left(\frac{\sum_t \mathbf{X}^2_{b,\,speech}(t)}{\sum_t \left(\mathbf{X}_{b,\,speech}(t) - \widehat{\mathbf{X}}_{b,\,speech}(t) \right)^2} \right) \tag{22}$$

where $X_{b,speech}$ (t) is the frequency-domain representation of the clean speech signal, for frequency b and time frame t, $\widehat{\mathbf{X}}_{b,\,speech}(t)$ is the frequency-domain representation of the estimated speech signal. PESQ indicates the quality difference between the enhanced and clean speech signals. PESQ is analogous to the mean opinion score, which is a subjective evaluation index. The PESQ score ranges from 0.5 to 4.5, and a high score indicates that the enhanced utterance is close to the clean utterance.

Contrary to spectral subtraction, dictionary approach does not assume a stationary interferer, optimizes the trade-off between source distortion and source confusion, and thus shows superiority over objective quality measures like cepstral distance, in the speaker-dependent and -independent case, in real-world environments and under low SNR condition. One possible reason could be due to lack of plenty of data to estimate a noise dictionary. At low SNR levels, the total volume of noise is much higher than that at high SNR levels, which offers a higher chance to obtain a good dictionary or noise modeling. However, under high SNR conditions, a lot of noise spectrum is buried in speech spectrum, which could make the learning of a noise dictionary difficult. The pretrained speech dictionary models outperform state-of-the-art methods like multiband spectral subtraction and approaches based on vector quantization [21–23]. Offline speech dictionary learning in a joint decomposition framework of the noisy speech spectrogram and a primary estimate of the clean speech spectrogram. Online learning approach processes input signals piece-by-piece by breaking the training data into

small pieces and updates learned patterns gradually using accumulated statistics. With this approach, only a limited segment of the input signal is processed at a time. The online estimated dictionary is sufficient enough in basis subspace to avoid speech distortion. The online approaches tend to give better performance than batch learning [53].

The computing demand for both offline learning and online learning consists of updating the coefficient matrix **C** and the pattern matrix **D**. The learning task is defined as an optimization problem, which aims to minimize an objective cost function $f(\mathbf{D})$ with respect to the pattern matrix **D**. It is observed that the reconstruction error for both the online and offline methods converges to a similar value after several iterations and not monotonically decreasing at the beginning. Both batch and online learning converge to a stationary point of the expected cost function $f(\mathbf{D})$ with unlimited data and unlimited computing resources. This situation is only valid in theory. For small-scale tasks where data are limited, but computing resources are unlimited, batch learning converges to a stationary point of the cost function $f_t(\mathbf{D})$, while online learning fails to converge, resulting in suboptimal patterns. For large-scale tasks, the more common situation is where training data are abundant but computing resources are limited. In this situation, due to its early learning property, online learning tends to obtain lower empirical cost than batch learning [49]. For sparse coding where the pattern matrix is overcomplete, for example, ($K > M$), then online learning is slower than batch learning. The online learning is significantly faster than the batch alternating learning by a factor of the large number of spectrograms reconstructed at each iteration [60].

In short, dictionary learning plays an important role in machine learning, where data vectors are modeled as sparse linear combinations of basis factors (i.e., dictionary). However, how to conduct dictionary learning in noisy environment has not been well studied. In this chapter, we have reviewed speech enhancement techniques based on dictionary learning. The dictionary learning-based algorithms have gained a lot of attention due to their success in finding high-"quality" dictionary atoms (basis vectors) that best describe latent features of the underprocessed data. As a multivariate data analysis and dimensionality reduction technique, two relatively novel paradigms for dimensionality reduction and sparse representation, NMF and SR, have been in the ascendant since its inception. They enhance learning and data representation due to their parts-based and sparse representation from the nonnegativity or purely additive constraint. NMF and SR produce high-quality enhancement results when the dictionaries for different sources are sufficiently distinct. This survey chapter mainly focuses on the theoretical research into dictionary learning-based speech enhancement where the principles, basic models, properties, algorithms, and employing on SR and NMF are summarized systematically.

Acknowledgements

This research is partially supported by the Ministry of Science and Technology under Grant Number 108-2634-F-008 -004 through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

Author details

Viet-Hang Duong[1], Manh-Quan Bui[2] and Jia-Ching Wang[2,3]*

*Address all correspondence to: jiacwang@gmail.com

1 Faculty of Information Technology, BacLieu University, Vietnam

2 Department of Computer Science Information Engineering, National Central University, Taiwan

3 Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan

References

[1] Loizou PC. Speech Enhancement: Theory and Practice. 1st ed. BocaRaton, FL: CRC Press; 2007

[2] Rabiner LR, Schafer RW. Theory and Application of Digital Speech Processing; 2001

[3] Gold B, Morgan N, Ellis D. Speech and Audio Signal Processing: Processing and Perception of Speech and Music. Berkeley, California, USA: Wiley; 2011

[4] Loizou PC. Speech Enhancement: Theory and Practice. Taylor and Francis; 2007

[5] Boll SF. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1979;**ASSP-27**(2):113-120

[6] Lu Y, Loizou PC. A geometric approach to spectral subtraction. Speech Communication. 2008;**50**:453-466

[7] Lim JS, Oppenheim AV. Enhancement and bandwidth compression of noisy speech. Proceedings of the IEEE. 1979;**67**(12):1586-1604

[8] Grancharov V, Samuelsson J, Kleijn B. On causal algorithms for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing. 2006;**14**(3):764-773

[9] Ephraim Y, Malah D. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Transactions on Audio, Speech, and Language Processing. 1984;**32**(6):1109-1121

[10] Martin R. Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors. IEEE Transactions on Audio, Speech, and Language Processing. 2005;**13**(5):845-856

[11] Cohen I. Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models. Signal Processing. 2006;**86**(4):698-709

[12] Erkelens JS, Hendriks RC, Heusdens R, Jensen J. Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors. IEEE Transactions on Audio, Speech, and Language Processing. 2007;**15**(6):1741-1752

[13] Chen B, Loizou PC. A Laplacian-based MMSE estimator for speech enhancement. Speech Communication. 2007;**49**(2):134-143

[14] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing. 2001;**9**(5):504-512

[15] Cohen I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Transactions on Speech and Audio Processing. 2003; **11**(5):466-475

[16] Hendriks RC, Heusdens R, Jensen J. MMSE based noise PSD tracking with low complexity. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Process (ICASSP). 2010. pp. 4266-4269

[17] Elad M, Aharon M. Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing. 2006;**15**(12):3736-3745

[18] Jiang Z, Lin Z, Davis LS. Label consistent K-SVD: Learning a discriminative dictionary for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;**35**(11): 2651-2664

[19] Huang J, Zhang T, Metaxas D. Learning with structured sparsity. Journal of Machine Learning Research. 2011;**12**:3371-3412

[20] Yaghoobi M, Blumensath T, Davies ME. Dictionary learning for sparse approximations with the majorization method. IEEE Transactions on Signal Processing. 2009;**57**(6): 2178-2191

[21] Sigg CD, Dikk T, Buhmann JM. Speech enhancement using generative dictionary learning. IEEE Transactions on Audio, Speech, and Language Processing. 2012;**20**(6):1698-1711

[22] Sun P, Qin J. Low-rank and sparsity analysis applied to speech enhancement via online estimated dictionary. IEEE Signal Processing Letters. 2016;**23**(12):1862-1866

[23] Sun M, Li Y, Gemmeke JF, Zhang X. Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence. IEEE Transactions on Audio, Speech, and Language Processing. 2015;**23**(7):1233-1242

[24] Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Transactions on Audio, Speech and Language Processing. 2013;**21**(10):2140-2151

[25] Chen Z, Ellis DP. Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2013. pp. 1-4

[26] Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research. 1997;**37**(23):3311-3325

[27] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;**401**(21):127-136

[28] Mallat S. A Wavelet Tour of Signal Processing—The Sparse Way. Academic Press; 2009

[29] Pauca VP, Piper J, Plemmons RJ. Nonnegative matrix factorization for spectral data analysis. Linear Algebra and its Applications. 2006;**416**(1):29-47

[30] Miao L, Qi H. End member extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. IEEE Transactions on Geoscience and Remote Sensing. 2007;**45**:765-777

[31] Li S, Hou X, Zhang H, Cheng Q. Learning spatially localized, parts-based representation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01). 2001. pp. 207-212

[32] Virtanen T. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on ASLP. 2007;**15**(3):1066-1074

[33] Ozerov A, Fevotte C. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Transactions on ASLP. 2010;**1**(3):550-563

[34] Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing. 2006; **54**(11):4311-4322

[35] Mavaddaty S, Ahadi SM, Seyedin S. Modified coherence-based dictionary learning method for speech enhancement. IET Signal Processing. 2015;**9**(7):537-545

[36] Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics. 1994;**5**:112-126

[37] Lee DD, Seung HS. Algorithms for nonnegative matrix factorization. In: Advances in Neural Information Processing Systems 13 (NIPS); 2000

[38] Cemgil AT. Bayesian inference for nonnegative matrix factorisation models. Computational Intelligence and Neuroscience. 2009;**2009**

[39] Févotte C, Bertin N, Durrieu JL. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural Computation. 2009;**21**:793-830

[40] Gonzales EF, Zhang Y. Accelerating the Lee-Seung Algorithm for Non-negative Matrix Factorization. Technical Report. Department of Computational and Applied Mathematics. Rice University; 2005

[41] Lin CJ. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Transactions on Neural Networks. 2007;**18**(6):1589-1596

[42] Lin CJ. Projected gradient methods for non-negative matrix factorization. Neural Computation. 2007;**19**:2756-2779

[43] Peharz R, Pernkopf F. Sparse nonnegative matrix factorization with ℓ0-constraints. Neurocomputing. 2012;**80**:38-46

[44] Eggert J, Korner E. Sparse coding and NMF. In: Proceedings of the 4th IEEE International Joint Conference on Neural Networks. 2004. pp. 2529-2533

[45] Schmidt MN. Speech Separation Using Non-negative Features and Sparse Nonnegative Matrix Factorization. Technical Report, Informatics and Mathematical Modelling. Technical University of Denmark, DTU; 2007

[46] Mairal J, Bach F, Ponce J, Sapiro G. Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research. 2010;**11**:19-60

[47] Lefevre A, Bach F, Fevotte C. Online algorithms for non-negative matrix factorization with the Itakura-Saito divergence. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA); Mohonk, NY. 2011

[48] Guan N, Tao D, Luo Z, Yuan B. Online nonnegative matrix factorization with robust stochastic approximation. IEEE Transactions on Neural Networks and Learning Systems. 2012;**23**(7):1087-1099

[49] Wang D, Vipperla R, Evans N. Online nonnegative convolutive pattern learning for speech signals. IEEE Transactions on Signal Processing. 2013;**61**(1)

[50] Yousefi M, Savoji MH. Supervised speech enhancement using online group-sparse convolutive NMF. In: Proceedings of the 8th International Symposium on Telecommunication (IST). 2016. pp. 494-499

[51] Robert T. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B (Methodological). 1996;**58**(1):267-288

[52] Kwon K, Shin JW, Kim NS. NMF-based speech enhancement using based update. IEEE Signal Processing Letter. 2015;**22**(4):44-56

[53] Rangachari S, Loizou PC. A noise-estimation algorithm for highly non-stationary environments. Speech Communication. 2006;**48**:220-231

[54] Varga A, Steeneken HJM, Tomlinson M, Jones D. The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical Report. Malvern, U.K.: DRA Speech Res. Unit; 1992

[55] http://www.dcs.shef.ac.uk/spandh/gridcorpus

[56] Vincent E, Gribonval R, Févotte C. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing. 2006;**14**(4): 1462-1469

[57] ITU-T Recommendation P.862. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. 2001

[58] Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICAASP). 2001:749-752

[59] Hu Y, Loizou PC. Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing. 2008;**16**(1):229-238

[60] Jose FR, Raviv R, Mauricio OA, Xiaoli ZF. Online learning of time-frequency patterns. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2017. pp. 2811-2815

[61] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallet DS, Dahlgren NL. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Philadelphia: Linguistic Data Consortium; 1993

IntechOpen