Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition

Author: Jort F. Gemmeke, Tuomas Virtanen and Antti Hurmalainen

Reporter: Yu-Hong Huang

Dept. of Computer Science and Engineering, National Sun Yat-sen University

21 Nov, 2011@NSYSU

Outline

- Abstract
- Introduction
- Model for noisy speech
- Sparse classification
- Sparse representation for feature enhancement
- Sparse representation for missing data technique
- Baseline recognizers
- Experiments

Abstract

- This paper proposes to use exemplar-based sparse representations for noise robust automatic speech recognition.
- First, we describe how speech can be modeled as a linear combination of a small number of exemplars from a large speech exemplar dictionary.
- The exemplars are time–frequency patches of real speech, each spanning multiple time frames.
- We then propose to model speech corrupted by additive noise as a linear combination of noise and speech exemplars, and we derive an algorithm for recovering this sparse linear combination of exemplars from the observed noisy speech.
- We describe how the framework can be used for doing hybrid exemplar-based/HMM recognition by using the exemplar-activations together with the phonetic information associated with the exemplars.

Abstract

- As an alternative to hybrid recognition, the framework also allows us to take a source separation approach which enables exemplar-based feature enhancement as well as missing data mask estimation.
- We evaluate the performance of these exemplar-based methods in connected digit recognition on the AURORA-2 database.
- Our results show that the hybrid system performed substantially better than source separation or missing data mask estimation at lower signal-to-noise ratios (SNRs), achieving up to 57.1% accuracy at SNR -5 dB.
- Although not as effective as two baseline recognizers at higher SNRs, the novel approach offers a promising direction of future research on exemplar-based ASR.

Introduction

- ASR using HMM+GMM for 30 years
- Background noise degrade performance
  - Reason
    - mismatch between training and testing data
  - Methods
    - normalization or enhancement of the features
    - Compensation of acoustic models
    - Use only the least noisy observation
    - …
  - Deal from stationary to non-stationary noise

Introduction

- Model based on *sparse representation*
  - Represent most information of a signal with linear combination of a small number of elementary signals, called *atoms*
  - Collection of atoms called a *dictionary*
- This paper investigate the effectiveness of combining two approaches

Introduction

- *Source separation*
  - Expressing a signal that is a mixture of multiple sources with sparse representation, using a dictionary for each underlying source
  - Finding the sparsest possible linear combination that describe the observed signal
  - Using techniques

    non-negative matrix factorization (NMF)

    Compressed sensing
  - Reconstruct using part of the dictionary pertaining single source
- Pattern recognition
  - Associating dictionary atoms with class labels
  - Using the weight of atoms in the sparse representation as evidence for the class of the observation signal
  - Lead the state-of-the-art classification result in various field

Introduction

  - We propose to use sparse classification in earlier work, in a hybrid SC/HMM speech recognizer
- We model signals as a sparse linear combinations of examples of that signal, then we model speech segments as a weighted linear combination of example speech segments, *exemplars*
  - These exemplars are spectrographic representations of speech spanning multiple time-frames of speech (50 to 300ms)
  - In traditional approach, speech is represented by one or more exemplars that each individually have the smallest distance to the observed speech token
  - In our framework, speech exemplars jointly approximate the observed speech

Introduction

- Dictionary using exemplars as atoms have several advantages
  - Relatively easy to construct by extraction of speech segments
  - Computationally efficient to construct dictionaries with high-dimensional atoms that contain several frames
    - Makes confusions between noise and speech atoms less likely
  - Allow very sparse representation if an observed speech segment closely resembles speech contained in the dictionary
  - The use of exemplar makes the mapping from atoms to speech classes straightforward
    - Each time-frame in the speech exemplars is directly labeled with an HMM-state label, obtained by means of a forced alignment using a conventional HMM-based recognizer

Introduction

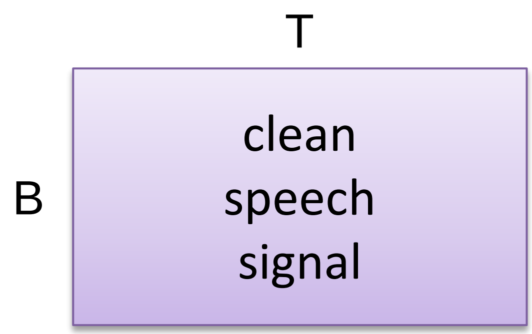
- In sparse classification approach, the weights of the linear combination of speech exemplars are used to provide a weighted sum of HMM-state scores for each frame in the observed speech
- In order to investigate the effectiveness of the sparse classification approach, we also use the exemplar-based sparse representations to apply two conventional robust ASR techniques
  - Feature enhancement, aims at providing clean speech features
  - Apply a missing data technique, for distinguished reliable or unreliable data, and discard the unreliable data, do imputation or marginalization of the missing feature

Introduction

- Contribution of this work are twofold
  - Investigate the effectiveness of combining two technique
  - Investigate to what extent using dictionary atoms that span multiple frames is beneficial for sparse representation- based noise robustness techniques
- Experiment compare the recognition accuracies of the various approaches using material from the AURORA-2 database
- Compare the recognition accuracy as a function of exemplar size

Model for noisy speech

- Sparse representation of noisy speech
  - Represent speech signal by *spectrogram*
  - Use the magnitude values directly
  - The magnitude spectrogram describe clean speech signal as a $B \times T$ dimensional matrix $\mathbf{S}$
    - B : frequency bands
    - T : time frames



  - The columns of this matrix are stacked into a single vector $\mathbf{s}$ of length $E = B \cdot T$, so that the entry $S(b,t)$, with $1 \leq b \leq B$ and $1 \leq t \leq T$, corresponding to the entry $s(b + (t-1)B)$

Model for noisy speech

- Assume that arbitrary speech spectrogram $\mathbf{s}$ can be expressed as a linear, non-negative combination of clean speech exemplars $\mathbf{a}_j^s$, with $j = 1, \dots, J$ denoting the exemplar index
- $\mathbf{s} \approx \sum_{j=1}^{J} \mathbf{a}_j^s x_j^s = \mathbf{A}^s \mathbf{x}^s$ subject to $\mathbf{x}^s \geq 0$
  - x_j^s : non-negative weight of each exemplar, or called *activation*
  - s *:* denote speech
  - $\mathbf{A}^s = [\mathbf{a}_1^s \mathbf{a}_2^s \dots \mathbf{a}_J^s]$ $(E \times J)$
  - $\mathbf{x}^s$ is a J-dimensional vector $(J \times 1)$
- $\mathbf{x}^s$ was shown to be sparse in previous research
- Noise spectrogram $\mathbf{N}$ can be represent by $\mathbf{n}$ as the linear combination of K noise exemplars $\mathbf{a}_k^n$, with $k = 1, \dots, K$ being the noise exemplar index

Model for noisy speech

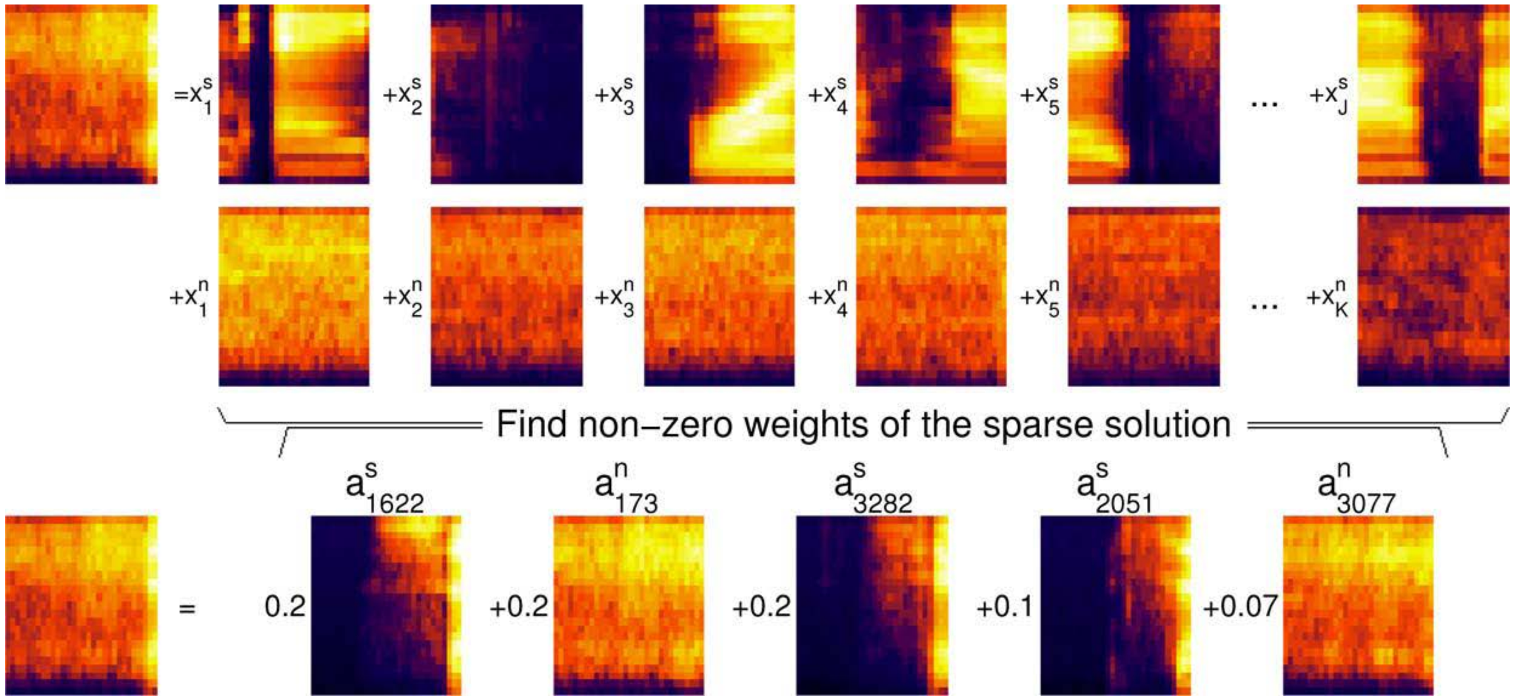
- Noisy speech segment $\mathbf{Y}$, reshaped into vector $\mathbf{y}$, can be a linear combination of both speech and noise exemplars

$$\mathbf{y} \approx \mathbf{s} + \mathbf{n} \approx \sum_{j=1}^{J} \mathbf{a}_j^s x_j^s + \sum_{k=1}^{K} \mathbf{a}_k^n x_k^n = [\mathbf{A}^{\mathrm{s}} \mathbf{A}^{\mathrm{n}}] \begin{bmatrix} \mathbf{x}^{\mathrm{s}} \\ \mathbf{x}^{\mathrm{n}} \end{bmatrix} = \mathbf{A}\mathbf{x}$$

  - $\mathbf{A}$: The whole speech and noise exemplar matrix $(E \times L), L = J + K$
  - $\mathbf{x}$: The activations of speech and noise exemplar $(L \times 1)$
- $\mathbf{x}$ is referred to as a sparse representation
- Normalize the dictionary rows and columns by iteratively scaling each row and column so that its Euclidean norm of column equals unity and row approximately equal
- During decoding, each noisy segment $\mathbf{y}$ is scaled using frequency band normalization applied to $\mathbf{A}$

Model for noisy speech

- Finding Activations
  - Represent the noisy speech $\mathbf{y}$ with model $\mathbf{Ax}$



Model for noisy speech

- The linear combination of exemplar is found by minimizing the cost function

$$d(\mathbf{y}, \mathbf{A}\mathbf{x}) + \|\boldsymbol{\lambda} .* \mathbf{x}\|_p$$

- The first term measure the distance between the noisy observation and the model
- Function d is the Kullback-Leibler (KL) divergence

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{e=1}^{E} y_e \log\left(\frac{y_e}{\hat{y}_e}\right) - y_e + \hat{y}_e$$

- In source separation method, the KL divergence has been found to produce better results than Euclidean distance

Model for noisy speech

- The second term is used in control the sparseness by the L_1 norm

$$\|\boldsymbol{\lambda}.*\mathbf{x}\|_1 = \sum_{l=1}^{L} x_l \lambda_l$$

- $\boldsymbol{\lambda}$ is used for penalizing all nonzero entries, this paper allow different weights for speech and noise exemplar in the dictionary
- Enforcing the sparseness of speech exemplar is very important

Model for noisy speech

- The cost function is minimized by first initializing the entries of the vector $\mathbf{x}$ to unity, and iteratively applying the update rule

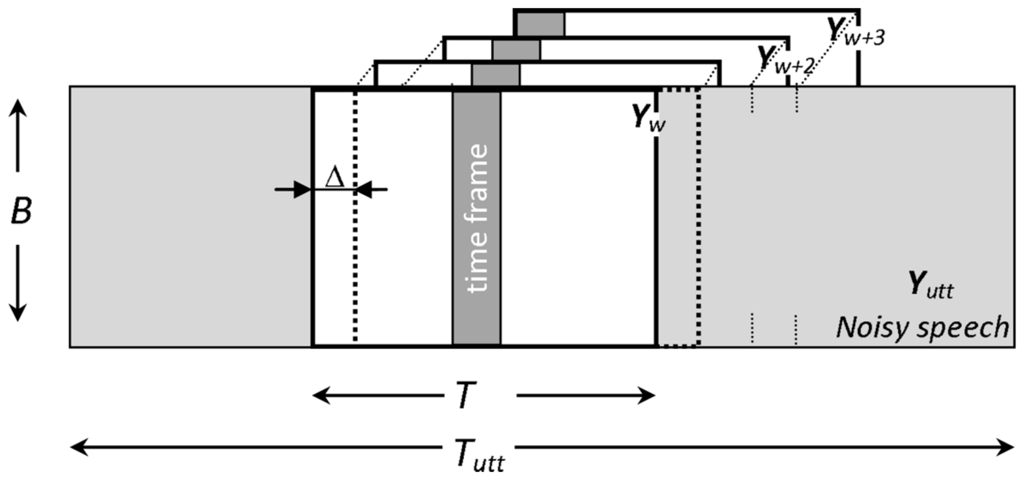
$$\mathbf{x} \leftarrow \mathbf{x}.* \frac{\left(\mathbf{A}^T\left(\frac{\mathbf{y}.}{\mathbf{Ax}}\right)\right).}{(\mathbf{A}^T\mathbf{1}+\boldsymbol{\lambda})}$$

- Where $\mathbf{1}$ is an all-one vector of length E
- $.*$ denoting element-wise multiplication, and so does division

- Sliding window approach for time continuity
  - In order to decode utterances of arbitrary length, a sliding time window approach was used
  - Dividing an utterance into a number of overlapping, fixed-length windows, the window length is equal to exemplar size T
  - We then find a sparse representation for each window

Model for noisy speech

- Consider noise speech utterance $\mathbf{Y}_{\text{utt}}$ represented as a magnitude spectrogram of size $B \times T_{\text{utt}}$
- Slide window, a matrix of size $B \times T$, through $\mathbf{Y}_{\text{utt}}$ using window shift Δ frame
- Obtain a sequence of windowed segments $\mathbf{Y}_1, \ldots, \mathbf{Y}_W$
- W is the number of windows in the utterance



Model for noisy speech

- Larger Δ reduce the computation effort, but decrease the accuracy
- We keep the window shift constant at $\Delta = 1$ frame
- At each window position w , the segment is reshape into observation vector $\boldsymbol{y}_w$
- The index w ranges from 1 to $W = T_{\text{utt}} - T + 1$
- The observation matrix $\boldsymbol{\Psi}$ of dimension $E \times W$ has the observation vector $\boldsymbol{y}_1 \ldots \boldsymbol{y}_w$ as its columns
- We can write $\boldsymbol{\Psi} \approx \mathbf{AX}$ s.t. $\mathbf{X} \geq 0$
- The column of activation matrix $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_W]$ consisting sparse representation of each window $(\mathrm{L} \times W)$

Sparse classification

- Preview
  - Sparse classification is a hybrid exemplar-based/HMM method
  - Keep the topology of HMM system
  - Rather than estimating the likelihoods of the states by means of GMMs, the calculation of likelihoods is based on the activations of exemplars
  - First introduce for the classification of isolated digit
  - Extended to enable the recognition of connected digit without noise
  - It can be used for noise robust connected digit recognition

Sparse classification

- Calculating speech state likelihood
  - Assuming a state-level labeling of each frame in speech data used to construct exemplar is available
  - Label each frame $t = 1, \dots, T$ in each exemplar $\mathbf{a}_j^s$ with a state label $q_{j,t} \in [1, Q]$ and form the label matrix $\mathcal{L}_j$, where Q is the total number of states
  - $\mathcal{L}_j$ is a sparse, binary matrix of dimensions $Q \times T$, the entries having values $[\mathcal{L}_j]_{q,t} = \delta(q, q_{j,t})$
  - δ is the Kronecher delta function

Sparse classification

- Denoting speech exemplar weights calculated for window w by $\mathrm{x}_{w,j}^{s}, j = 1, \ldots, J$, we calculate state likelihood matrix

$$\boldsymbol{L}_w = \sum_{j=1}^{J} \mathcal{L}_j \, \mathrm{x}_{w,j}^{s}$$

- The columns of $\boldsymbol{L}_w$ are denoted with vector $\boldsymbol{l}_{w,t}, t = 1, \ldots, T$
- Overlapping windows are combined by summing the likelihood of the frames of all windows in each they occur
- The combined state likelihood vector $\mathbf{l}_\tau^{\mathrm{utt}}$ for each frame $\tau = 1, \ldots, T_{\mathrm{utt}}$ is given as

$$\mathbf{l}_\tau^{\mathrm{utt}} = \sum_{t=\max(1,\tau-T_{\mathrm{utt}}+T)}^{\min(T,\tau)} \mathbf{l}_{\tau-t+1,t}$$

Sparse classification

- After obtaining the state likelihoods for entire utterance, we use the Viterbi algorithm to find the state sequence that maximizes total likelihood

- Silence likelihoods
  - The likelihood of silence cannot be reliably estimated from noisy utterances
    - Silence is absence of speech energy, a sparse representation of magnitude spectrograms models silence with all exemplar weights close or equal to zero
    - The state likelihoods are calculated by multiplication of the atom activations with the label matrix, and the silence state likelihood will be very low, and will have numerous insertion errors
  - Modify the speech and silence likelihood
    - Measure the activity of speech and noise exemplar
    - Boosting the silence likelihood when there is no speech activity

Sparse representation for FE

- We use the sparse representations of speech and noise to estimate clean speech spectrograms, i.e., do feature enhancement
- Denoting the spectrum vector of tth frame of speech exemplar j by $\mathrm{a}_{j,t}^{s}$, the clean speech estimate $\tilde{s}$ for the tth frame of window w can be written as

$$\tilde{\boldsymbol{s}}_{w,t} = \sum_{j=1}^{J} \mathbf{a}_{j,t}^{s} x_{j,w}^{s}$$

- And noise estimate $\tilde{n}$ is given by

$$\tilde{\boldsymbol{n}}_{w,t} = \sum_{k=1}^{K} \mathbf{a}_{k,t}^{n} x_{k,w}^{n}$$

Sparse representation for FE

- For each frame $\tau = 1, \ldots, T_{\mathrm{utt}}$ of the utterance, the model pertaining to overlapping windows are summed to obtain the speech and noise models

$$\hat{\boldsymbol{s}}_{\tau} = \sum_{t=\max(1,\tau-T_{\mathrm{utt}}+T)}^{\min(T,\tau)} \tilde{\boldsymbol{s}}_{\tau-t+1,\tau}$$

$$\hat{\boldsymbol{n}}_{\tau} = \sum_{t=\max(1,\tau-T_{\mathrm{utt}}+T)}^{\min(T,\tau)} \tilde{\boldsymbol{n}}_{\tau-t+1,\tau}$$

- The resulting frame-wise estimates are grouped into speech and noise spectrogram utterance matrices

$$\hat{\boldsymbol{S}}_{\mathrm{utt}} = \left[\hat{\boldsymbol{s}}_1, \ldots, \hat{\boldsymbol{s}}_{T_{\mathrm{utt}}}\right]$$

$$\hat{\boldsymbol{N}}_{\mathrm{utt}} = \left[\hat{\boldsymbol{n}}_1, \ldots, \hat{\boldsymbol{n}}_{T_{\mathrm{utt}}}\right]$$

Sparse representation for FE

- The reconstructed speech spectra could be used directly as an estimate of clean speech features
- We obtain better results by using a time-varying filter

$$\boldsymbol{h}_t = \frac{\hat{\boldsymbol{s}}_{\mathrm{t}}}{(\hat{\boldsymbol{s}}_{\mathrm{t}} + \hat{\boldsymbol{n}}_{\mathrm{t}})}$$

- Calculate the enhancement features in each frame as

$$\boldsymbol{h}_t .* \boldsymbol{y}_t$$

Sparse representation for MDT

- Missing data technique is known for its high accuracy at high SNRs and its ability for dealing non-stationary noise type
- MDT can estimate which spectro-temporal elements in the spectrogram are reliable (dominated by speech) or unreliable (dominated by noise)
- The reliability estimate of noisy speech features are referred to as a *missing data mask*

$$M_{\text{utt}}(b,\tau) = \begin{cases} 1 = \text{reliable}, & \text{if } \dfrac{\hat{S}_{\text{utt}}(b,\tau)}{\widehat{N}_{\text{utt}}(b,\tau)} > \theta \\ 0 = \text{unreliable}, & \text{otherwise} \end{cases}$$

- The constant θ is an empirically determined SNR threshold

Baseline recognizers

- Compare the results obtained with the exemplar-based framework with two noise robust recognizer
  - Multi-condition trained recognizer with mean and variance normalization to achieve state-of-the-art performance
  - MDT-based recognizer employing so-called harmonicity missing data mask
    - In the harmonicity mask, the noisy speech signal is first decomposed in a harmonic and a residual part using a least squares fitting method
    - The harmonic energy is then used as an estimator for the noisy energy

Experiments

- Use AURORA-2 database with the five methods
- Experiment setup
  - Recognition task
    - Test set A and B
    - Training material: clean and multi-condition data set, each containing 8440 utterances
  - Finding sparse representations
    - Mel frequency magnitude spectra B = 23 frequency band
    - Center frequency starting at 100Hz
    - Hamming window with frame length 25ms and frame shift 10ms
    - Exemplar based framework was implemented in MATLAB
    - Update rule was run for 200 iterations and converged
    - $\lambda = 0.65$ for speech exemplars and $\lambda = 0$ for noise exemplars

Experiments

– Dictionary creation
  - For four kind of exemplar size $T \in \{5, 10, 20, 30\}$ frames
  - Select two segments of length T with random offsets for every utterances in multi-condition training set
  - The underlying clean speech and noise were extracted and added to the speech and noise dictionaries
  - Randomly select 4000 exemplars from speech dictionary
  - Remove silence, randomly select 4000 exemplars from noise dictionary
  - The choice of random subset did not influence recognition significantly

– Speech recognizers
  - Multi-condition trained recognizer is the HTK-based recognizer
    - use c0 in place of log energy

Experiments

- All other experiments use the MATLAB implementation of the HMM-based missing data recognizer
- The acoustic models, trained on clean speech in the training set consist
  - 11 whole-word models with 16 states
  - $Q = 179$ dimensional state-space (16*11+3)
  - Each state was modeled by a mixture of 16 Gaussians with diagonal covariance
- MDT
  - Missing data mask used for missing data baseline is harmonicity mask with $10\log_{10}\theta = -9$ dB
  - Missing data mask provided by exemplar-based framework, $10\log_{10}\theta = \{-2, -1, 0, 0\}$ dB for exemplar size $\{5, 10, 20, 30\}$
- FE: Single-pass retraining and re-estimating
- SC: Force alignment of clean speech training set, used for labeling the speech dictionary

Experiments

- Experiment results
  - M: Multi-condition
  - I: MDT baseline
  - SMDT: sparse missing data
  - FE: feature enhancement
  - SC: sparse classification
  - SNR: inf, 15, 5, -5
  - T: exemplar size 5, 10, 20, 30