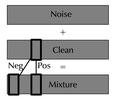


Learning a concatenative resynthesis system for noise suppression

Michael I Mandel, Young Suk Cho, and Yuxuan Wang · Computer Science & Engineering · The Ohio State University

THE OHIO STATE UNIVERSITY

1. Motivation

Scenario your phone records your voice, detects clean segments, replaces noisy segments with "matching" pre-recorded clean segments

Learned data-driven matching function

Non-linear use psychoacoustically motivated features in matching

Dictionary-based speech model improves as clean recordings accumulate

Concatenative transition model creates full concatenative synthesizer

2. Contributions

- Analysis-by-synthesis approach to non-parametric noise suppression
- Deep neural network (DNN) matching for concatenative resynthesis
- Speaker-dependent experiments on small-vocab CHiME2-GRID corpus show very high quality, but slightly compromised intelligibility

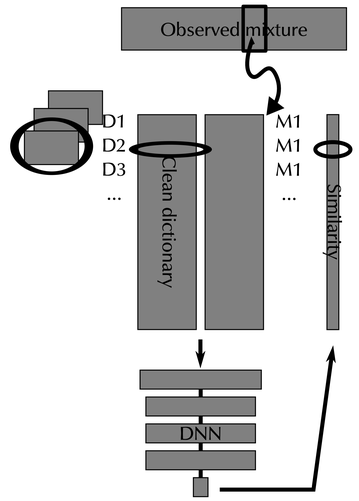
3. Benefits

- Replaces noisy speech with noise-free synthetic speech
  vs. mask-based noise suppression, (e.g., Narayanan and Wang, 2014)
- Uses a rich dictionary of speech signals
  vs. parametric systems like mask predictors
- Uses perceptually relevant non-linear audio features
  vs. linear methods like non-negative matrix factorization (Lee and Seung, 1999) and sparse coding (Plumbley et al., 2010)
- Does not need to be trained on specific test-time noises
  vs. existing dictionary-based noise suppression systems (Xiao and Nickel, 2010; Ming et al., 2011; Ogawa et al., 2014)

4. Future Directions

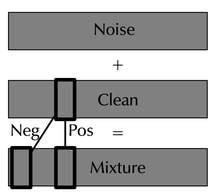
- Use for bandwidth expansion, audio super-resolution, dereverberation
- Apply to Large-vocabulary tasks
- Pre-analyze dictionary for phonetic and/or pitch information
  - Allows non-parametric, noise robust ASR and pitch tracking
- Match noise-suppressed speech to clean dictionary like Delcroix et al. (2011); Nickel et al. (2013)
- Use more sophisticated transition (language) model
- Apply to musical source separation

5. System overview



Each "chunk" of the observed mixture signal is paired with every "chunk" of clean speech in the dictionary and the pair is input to a deep neural network (DNN), providing a similarity score. The noisy chunk sequence is replaced by the best sequence of dictionary chunks according to this score and a transition affinity between dictionary chunks.

6. Training data: clean-noisy chunk pairs



Two pairs of clean and noisy chunks illustrating the collection of training data for the paired-input DNN. In the positive pair, the clean speech chunk is present in the noisy mixture chunk; in the negative pair it is not.

7. Mathematical details

- $\hat{\mathbf{z}}$: Best sequence of dictionary chunks for noisy sequence $\{x_t\}_{t=0}^{T}$

$$\hat{\mathbf{z}} = \operatorname*{argmax}_{\mathbf{z}} \prod_t p(z_t = j \mid x_t) p(z_t = j \mid z_{t-1} = i) \quad (1)$$

$$= \operatorname*{argmax}_{\mathbf{z}} \prod_i g(z_j, x_i) T_{ij} \quad (2)$$

- $g(z_j, x_i)$: similarity function between clean chunk z_j and noisy chunk x_i
  - trained by minimizing cross-entropy to y_{ij}

$$\sum_{i,j} -y_{ij} \log g(z_j, x_i) - (1 - y_{ij}) \log(1 - g(z_j, x_i)). \quad (3)$$

  - y_{ij}: 1 if noisy chunk z_j contains clean chunk x_i, else 0
- T_{ij}: transition matrix from clean chunk z_i to z_j

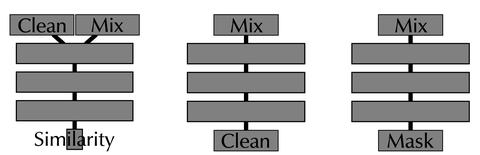
$$T_{ij} = \exp\left(-d_\tau(z_i, z_j)/\gamma\right) \quad (4)$$

  - $d_\tau(z_i, z_j)$: Euclidean distance between log mel spectra of the last τ frames of z_i and the first τ frames of z_j
  - γ: parameter controlling mapping of distances to affinities

8. Small vocab enhancement experiments

Dataset Small-vocabulary CHiME2-GRID corpus, speaker 3 (male)
Speech sentences of the form "⟨*command*⟩ ⟨*color*⟩ ⟨*preposition*⟩ ⟨*letter*⟩ ⟨*digit*⟩ ⟨*adverb*⟩", e.g.,"place blue at A 9 again."
Noise household noise, SNR of −6 to 9 dB
Features log mel spectrum of 196 ms "chunks"
Train on "isolated" noisy train set (500 sentences) to map noisy reverberant speech to reverberant speech (124,080 clean-noisy chunk pairs)
Test on isolated noisy devel set, random SNR for each of 24 mixtures
Dictionary reverberant speech from full training set (67,040 chunks)
Comparison to mask-based approaches trained on the same data
Evaluation subjective quality and intelligibility

9. DNNs compared



(a) Proposed (b) Predict clean (c) Predict mask

Structure of the DNNs used in the experiments. Predicting (a) similarity from clean-noisy chunk pair, (b) clean chunk from noisy chunk, (c) ideal ratio mask from noisy chunk.

10. Online audio examples

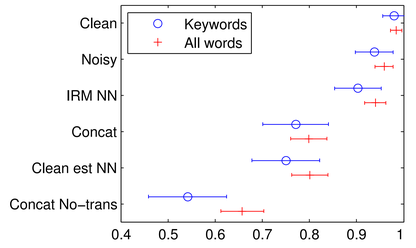
http://mr-pc.org/work/globalsip14/

11. Preliminary: chunk retrieval

- Idealized experiment: test retrieval performance when dictionary contains exact correct clean chunk
- Averaged over 500 test chunks from CHiME2-GRID devel set
- For average rank, lower is better

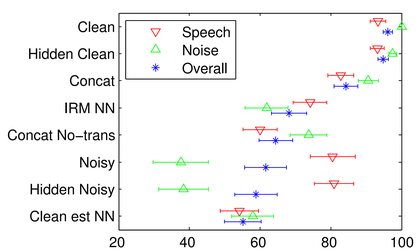
	Euclidean	Paired NN
Number of dictionary chunks	2899	2899
Average rank of correct chunk	630	5
Precision-at-1	21.5%	72.3%

12. Subjective intelligibility results



Intelligibility subjective listening test results showing the average percentage correctly identified of all words and just the keywords (letters and numbers) for each system. Errors bars show 95% confidence intervals.

13. Subjective quality results



MUSHRA listening test results showing the average subjective speech quality, noise suppression, and overall quality for each system. Higher numbers are better in all cases. Errors bars show 95% confidence intervals.

References

Delcroix, M. et al. (2011). "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation". In: *Proc. 1st Intl. Workshop on Machine Listening in Multisource Environments (CHiME)*, pp. 12–17.
Lee, D. D. and H. S. Seung (1999). "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755, pp. 788–791.
Ming, J., R. Srinivasan, and D. Crookes (2011). "A Corpus-Based Approach to Speech Enhancement From Nonstationary Noise". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 19.4, pp. 822–836.
Narayanan, A. and D. Wang (2014). "Investigation of speech separation as a front-end for noise robust speech recognition". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 22, pp. 826–835.
Nickel, R. M., R. F. Astudillo, D. Kolossa, and R. Martin (2013). "Corpus-Based Speech Enhancement With Uncertainty Modeling and Cepstral Smoothing". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 21.5, pp. 983–997.
Ogawa, A., K. Kinoshita, T. Hori, T. Nakatani, and A. Nakamura (2014). "Fast segment search for corpus-based speech enhancement based on speech recognition technology". In: *Proc. IEEE Intl. Conf. Acous. Speech & Sig. Proc.*
Plumbley, M. D., T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies (2010). "Sparse Representations in Audio and Music: From Coding to Source Separation". In: *Proc. IEEE* 98.6, pp. 995–1005.
Xiao, X. and R. M. Nickel (2010). "Speech Enhancement With Inventory Style Speech Resynthesis". In: *IEEE Tr. Aud. Speech. Lang. Proc.* 18.6, pp. 1243–1257.