Concatenative Resynthesis with Improved Training Signals for Speech Enhancement

Ali Raza Syed[1], Trinh Viet Anh[1], Michael I Mandel[2]

[1] The Graduate Center, City University of New York, NY – USA , [2] Brooklyn College, City University of New York, NY, USA

1. Concatenative Resynthesis

What is it? Methods that replace short units of noisy speech with matching units of clean speech, thus producing a noiseless signal while improving the quality of the speech (Mandel 2014).

Core component. Use a DNN to learn a *non-linear similarity metric* between short units of noisy and clean speech.

Current limitation. *Exact match* criterion is required to train the DNN.

Given:

- chunks of clean speech $\{z\}_{j=1}^{J}$
- chunks of noisy speech $\{x\}_{i=1}^{I}$
  - where $x_i = z_i \oplus \text{noise}$

Exact match criterion uses paired examples for training:

- positive examples: (z_i, x_i)
- negative examples: (z_j, x_i) with $i \neq j$

What's the problem? This restrictive criterion limits the training signal available for learning to the DNN.

Goal

Improve the training signal available to the DNN to learn a more useful and generalizable similarity function.

Proposed approach: Rather than exactly matching clean and noisy examples, select pairs that are *sufficiently similar using* acoustic, phonetic and prosodic characteristics of the signal. Thus, make more efficient use of the available data by allowing substitution of a clean signal with another *similar* clean signal.

2. Technical Overview

Data

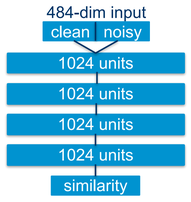
- CHiME2-GRID small vocabulary dataset with utterances using fixed grammar: "*command color preposition letter digit adverb.*"
- Household noises at different SNR
- Single speaker
- Train set: 460 utterances
- Validation set: 40 utterances
- Test set: 24 utterances

Feature extraction

- Compute log mel spectrograms
- Extract 11-frame *chunks* of duration 192 ms
- Each chucks overlaps with its neighbors by 10 frames
- Training set: 113,896 paired examples

Paired-input network (DNN)

- input: concatenated clean & noisy vectors
- 4 hidden layers
- output: 2-unit softmax layer
- cross-entropy loss
- Adam SGD



Baseline

- DNN with exact match criterion from Mandel et al., 2014 [1]

3. Methods

3a. Phonetic Similarity

Find pairs of chunks $(z_i, x_j), i \neq j$ that are sufficiently similar phonetically to be substitutes and thus serve as positive examples for the DNN.

Frame-wise phonetic similarity:

$$s_{Ph}(z_i, z_j) = \frac{1}{F}\sum_{k=1}^{F}\delta(p_k^{(z_i)}, p_k^{(z_j)})$$

where δ is the Kronecker *delta* function such that $\delta(u, v) = 1$ when $u = v$ and $\delta(u, v) = 0$ otherwise.

In our case:

- $F = 11$ frames per chunk
- A chunk is labeled as $z_i = p_1^{(z_i)} \cdots p_F^{(z_i)}$ with phonetic labels $p_k \in [1, 38]$
- threshold for positive examples: $s_{Ph} \geq \frac{8}{11}$
- threshold for negative examples: $s_{Ph} \leq \frac{3}{11}$
- resulted in a training set: ~1,500,000 paired examples (tenfold increase from previous work [2])

3b. Perceptual Similarity

Improve the discriminative power of the DNN for distinguishing commonly confused phones in speech.

- Use perceptual confusion to categorize the phonetic labels into 10 groups: *stressed vowels, unstressed vowels, voiced plosives, unvoiced plosives, affricates, voiced fricatives, unvoiced fricatives, approximants, nasals, and silence*.
- Represent a chunk by a vector $q^{(z_i)} = \{q_1^{(z_i)}, \ldots, q_F^{(z_i)}\}$ where $q_k \in \{1 \ldots 10\}$ represents one the ten groups.

Frame-wise perceptual similarity:

$$s_Q(z_i, z_j) = \frac{1}{F}\sum_{k=1}^{F}\delta(q_k^{(z_i)}, q_k^{(z_j)})$$

In our case:

- threshold for negative examples: $s_q \geq \frac{8}{11}$ while phonetically dissimilar with $s_{Ph} \leq \frac{8}{11}$

3c. Acoustic and Prosodic Similarity

Use per-frame acoustic intensity, fundamental frequency, and periodicity to evaluate the similarity between chunks.

Frame-wise intensity similarity:

$$s_I(I_i, I_j) = e^{-\alpha_I \|I_i - I_j\|}$$

Use similar approach for fundamental frequency (with α_F) and periodicity (with α_{Pe}).

Overall frame-wise acoustic and prosodic similarity:

$$s(z_i, z_j) = \lambda_{Ph} s_{Ph}(z_i, z_j) + \lambda_I s_I(z_i, z_j) + \lambda_F s_F(z_i, z_j) + \lambda_{Pe} s_{Pe}(z_i, z_j)$$

In our case:

- values for hyperparameters $\lambda_{Ph}, \lambda_I, \lambda_F, \lambda_{Pe}, \alpha_I, \alpha_F, \alpha_{Pe}$: Bayesian optimization minimizing the frame-wise error rate on the development set

4. Evaluation

4a. Objective

To evaluate the quality of the mapping, $x \to z$, we can compare the frame-level phonetic transcriptions of x and z.

Frame-wise error rate:

$$e_f(z, x) = 1 - \frac{1}{F}\sum_{k=1}^{F}\delta(p_k^{(z)}, p_k^{(x)}), (2)$$

Phone error rate: score with SCTK after collapsing repeated phones and aligning the phonetic sequences of the clean and denoised signals. Reconstructed chunks overlap by 5 frames; overlapping frames with differing labels are annotated with composite phones.

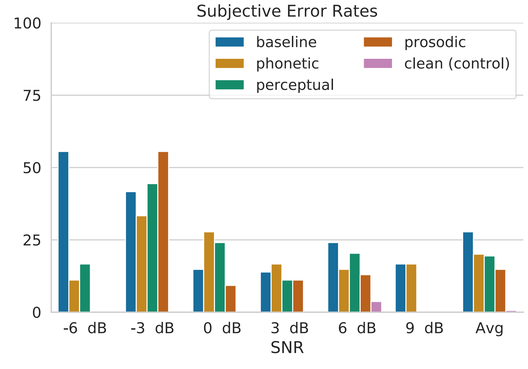
4b. Subjective

Intelligibility listening test: within-subject design with 3 participants.

Stimuli: 60 utterances based on 12 randomly selected noisy utterances from the test set

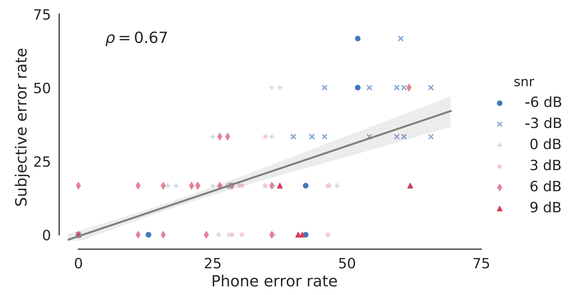
- 12 resynthesized using baseline
- 12 resynthesized using phonetic similarity
- 12 resynthesized using perceptual similarity
- 12 resynthesized using acoustic and prosodic similarity
- 12 clean counterparts serving as control

Subjective intelligibility error rate: proportion of words incorrectly transcribed by the participants.



5. Objective versus Subjective Metrics

Subjective intelligibility error rates were found to be strongly correlated with the objective frame-wise and phone error rates.



6. Summary

- Combining similarity based on phonetic content and periodicity of the signals yields noiseless speech that is more intelligible to humans.
- Computed error metrics can be effectively used for tuning and evaluating such systems.
- It is possible to perform this in an efficient and scalable manner by using approximate nearest neighbors for aiding in the selection of training data.
- There is a correlation of computed error metrics with subjective human evaluations based on intelligibility listening tests.

Future work: Investigate performance and scalability on large vocabulary and speaker independent speech.

7. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1618061. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. I. Mandel, Y. S. Cho, and Y. Wang, "Learning a concatenative resynthesis system for noise suppression," in 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Dec 2014, pp. 582–586.

[2] M. I. Mandel and Y. S. Cho, "Audio super-resolution using concatenative resynthesis," in Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on, 2015, pp. 1–5.