Interspeech 2018
2-6 September 2018, Hyderabad

Concatenative Resynthesis with Improved Training Signals for Speech Enhancement

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

[1] The Graduate Center, CUNY, New York, NY, USA
[2] Brooklyn College, CUNY, New York, NY, USA

asyed2@gradcenter.cuny.edu, vtrinh@gradcenter.cuny.edu, mim@sci.brooklyn.cuny.edu

Abstract

Noise reduction in speech signals remains an important area of research with potential for high impact in speech processing domains such as voice communication and hearing prostheses. We extend and demonstrate significant improvements to our previous work in synthesis-based speech enhancement, which performs concatenative resynthesis of speech signals for the production of noiseless, high quality speech. Concatenative resynthesis methods perform unit selection through learned non-linear similarity functions between short chunks of clean and noisy signals. These mappings are learned using deep neural networks (DNN) trained to predict high similarity for the exact chunk of speech that is contained within a chunk of noisy speech, and low similarity for all other pairings. We find here that more robust mappings can be learned with a more efficient use of the available data by selecting pairings that are not exact matches, but contain similar clean speech that matches the original in terms of acoustic, phonetic, and prosodic content. The resulting output is evaluated on the small vocabulary CHiME2-GRID corpus and outperforms our original baseline system in terms of intelligibility by combining phonetic similarity with similarity of acoustic intensity, fundamental frequency, and periodicity.

Index Terms: noise reduction, speech enhancement, speech processing, concatenative resynthesis, deep neural networks.

1. Introduction

The presence of noise in speech signals can significantly deteriorate the performance of speech processing applications in domains such as voice communication and hearing prostheses. Research in noise reduction largely targets modification of the noisy signal to approximate the clean signal. While such approaches improve the overall quality of the signal, they also tend to reduce the quality of the clean speech while retaining some noise [1]. Thus, there is interest in methods which can transform noisy speech to produce high quality and noiseless speech. Concatenative resynthesis methods [2, 3] replace short units of noisy speech with matching units of clean speech, thus producing a noiseless signal while improving the quality of the speech. The core component of these methods is to learn a non-linear similarity metric between short units of noisy and clean speech. Given a noisy speech signal and a dictionary of units of clean speech, these methods use the learned metric to identify the most similar clean units of speech that can be concatenated to reconstitute the original clean speech signal. The non-linear similarity metric is learned using a deep neural network (DNN) which is presented with paired examples of clean and noisy speech. In previous work [2, 3, 4], such pairings were based on an exact match criterion. Each paired training example consists of a clean signal and its noisy counterpart, *i.e.* where noise has been added to that clean signal. The DNN is trained to learn a similarity score of 1 when presented with such a positive example, and a score of 0 for any other pairing (negative example). To balance positive and negative classes, the negative pairs are sampled by associating a clean signal with a random noisy signal which is not its counterpart. This is a restrictive criterion and limits the training signal available for learning to the DNN. It constrains the amount of training data and does not allow for generalization across signals with similar, but not identical, acoustic, and linguistic content.

We investigate improving the training signal by taking into account the acoustic, phonetic and prosodic similarity of paired signals. Rather than exactly matching clean and noisy examples, we select pairs that are sufficiently similar based on one or more of these characteristics. Thus the DNN may learn to substitute a clean signal with another clean signal containing similar speech content. This makes more efficient use of the available data since we no longer need to rely on exact matches to select paired examples for training the DNN. Moreover, it allows for the DNN to learn a more generalizable similarity metric.

2. Related Work

Concatenative resysnthesis for producing noiseless, high quality speech from noisy signals was introduced by Mandel et al. [2, 3] and used a DNN for similarity metric learning when presented with paired examples (concatenated feature vectors from pairs of clean and noisy units). Maiti et al. [4] extended this work to use twin neural networks with ranking loss for learning the similarity metric. However, the neural network in both systems used paired examples that were selected using an exact match criterion. Our work uses acoustic and linguistic characteristics of the clean speech to select pairs that are sufficiently close to be substituted for one another. The system described by Mandel et al. [2] also serves as the baseline for comparison in our experiments.

Concatenative resynthesis systems are an example of exemplar-based speech enhancement. Exemplar-based methods have been proposed in recent research using generative models for clean and noisy signals. Xiao et al. [5] and Nickel et al. [6] used Gaussian mixture model hidden Markov model (GMM-HMM) based systems to perform coarse matchings against a dictionary of signals to select single best noise-suppressed exemplars. Ming et al. [7], Delcroix et al. [8], and Ogawa et al. [9], replaced subsequences of noisy utterances with sub-sequences of the training corpus having maximum likelihood under their GMM. By using generative models, these approaches are optimized to describe their training data. Concatenative resysnthesis using DNNs, as in our approach, differs by learning a similarity metric from training examples.

10.21437/Interspeech.2018-2439

3. Technical Overview

3.1. Data

We use the CHiME2-GRID small vocabulary dataset [10] of simulated speech recorded in a living room environment. Each utterance is a six-word sentence from the GRID corpus [11] of the form "*command color preposition letter digit adverb*", *e.g.* "place blue at F 9 now". The recordings are mixed with household noises at six different signal-to-noise ratios (SNR): -6, -3, 0, 3, 6, and 9 dB.

We use recordings from a single speaker (id 3) for training and testing with clean speech from the "reverberant" condition and noisy speech from the "isolated" condition (following the experiments of Mandel et al. [2]). Stereo signals are averaged to obtain monaural signals. The official training set has 500 utterances from which we randomly selected 40 utterances for a validation set and use the remaining 460 utterances for training. Our test set consisted of 24 utterances from the official development set, selected over six different SNR values. In all experiments, our training set was selected to balance the number of positive and negative examples of paired inputs.

3.2. Feature extraction

For each utterance, we compute log mel spectrograms with FFT frame size of 32 ms and hop size of 16 ms. From these, we extract 11-frame chunks of duration 192 ms, represented as 242-dimensional feature vectors. Each chunks overlaps its neighbors by 10 frames. This yields a training set of 113,896 examples of both clean and noisy speech. For resynthesis at test time, we used all 500 utterances to build a dictionary of clean speech chunks available for matching to noisy chunks.

3.3. Paired-input network

The core component in concatenative resynthesis is the paired-input network, a DNN for learning a non-linear similarity metric $g(z, x)$ between chunks of clean speech $\{z\}_{i=1}^{I}$ and chunks of noisy speech $\{x\}_{j=1}^{J}$. We assume a noisy chunk is a clean chunk superimposed with noise. A paired example (z_i, x_j) is represented as a 484-dimensional vector by concatenating clean and noisy feature vectors. This is input to a DNN with 4 hidden layers, each with 1,024 rectified linear units (ReLU). We denote the target metric with $y_{ij} \in \{0, 1\}$ with positive examples as 1 and negative examples as 0. The output is a 2-unit softmax layer representing the probability of belonging to a binary class $\{0, 1\}$, the target similarity metric. The DNN is trained to minimize cross-entropy loss, $\mathcal{L}(y_{ij}, g(z_i, x_j))$:

$$-\sum_{i,j} y_{ij} \log g(z_i, x_j) + (1 - y_{ij}) \log(1 - g(z_i, x_j))). \quad (1)$$

The DNN is trained using Adam stochastic gradient descent [12] with initial learning rate 0.015 and decay parameters 0.9 and 0.999 (first and second moments). The batch size is 512 and we use a dropout probability of 0.2 for hidden units [13].

4. Methods

In previous work [2], positive examples were selected using an exact-match criterion yielding pairs (z_i, x_i), while negative examples were randomly assigned per clean chunk yielding pairs $(z_i, x_j), i \neq j$. This strict criterion restricts the number of pairs available for training and limits the training signal available to the DNN by forgoing examples that are sufficiently similar to be substitutable. Thus, we investigate different methods of improving the training signal available to the DNN to learn a more useful and generalizable similarity function.

4.1. Phonetic similarity

Any collection of speech recordings contains many examples of short chunks which realize identical or similar phones. In this work, we leverage this fact, by also considering pairs of chunks that are sufficiently similar to be considered as substitutes. Such paired chunks provide positive examples of the form (z_i, x_j) where $i \neq j$ necessarily. Each frame in a chunk z_i can be annotated at the frame level by the phone realized during that interval of speech. Thus the chunk may also be represented as a phonetic vector, $p^{(z_i)} = \{p_1^{(z_i)}, \ldots, p_F^{(z_i)}\}$, where F is the number of frames per chunk and $p_k \in \{1 \ldots 38\}$ is an integer representing one of 38 possible phonetic labels. We perform forced alignment of utterances with their phonetic transcription using the Montreal Forced Aligner [14]. Using this alignment, each frame is annotated with a phone corresponding to its time interval. We compute the frame-wise phonetic similarity as:

$$s_{Ph}(z_i, z_j) = \frac{1}{F} \sum_{k=1}^{F} \delta(p_k^{(z_i)}, p_k^{(z_j)}), \quad (2)$$

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 and we consider two chunks to be sufficiently similar when $s_{Ph} \geq \frac{8}{11}$, *i.e.* we require a frame-wise phonetic correspondence of at least 8 frames. By finding such pairs, we build a dataset of positive examples. For negative examples, we apply a threshold of $s_{Ph} \leq \frac{3}{11}$ to ensure that paired chunks are sufficiently dissimilar. Our experiments used the approximate nearest neighbors library, NMSLIB [15], to efficiently find positive examples. Exploring our dataset of signals we found, unsurprisingly, that overwhelming number of pairs with zero phonetic similarity. Thus, we were able to efficiently select negative examples by randomly sampling pairs and pruning them to discard pairs with $s_{Ph} \leq \frac{3}{11}$. Using this approach, we were able to efficiently construct training sets with up to 1,500,000 paired examples. For comparison, the previous work [2], which serves as our baseline, was limited to the number of clean chunks, well under 150,000 training pairs.

4.2. Perceptual similarity

We also investigate improving the discriminative power of the DNN for distinguishing commonly confused phones in speech. We categorized the phonetic labels into ten groups based on classic results in perceptual confusion by Miller and Nicely [16]: stressed vowels, unstressed vowels, voiced plossives, unvoiced plossives, affricates, voiced fricatives, unvoiced fricatives, approximants, nasals, and silence. Each chunk was then represented by a vector $q^{(z_i)} = \{q_1^{(z_i)}, \ldots, q_F^{(z_i)}\}$ where $q_k \in \{1 \ldots 10\}$ representing one the ten groups. The perceptual similarity of two chunks can be computed in a similar way as the phonetic similarity:

$$s_Q(z_i, z_j) = \frac{1}{F} \sum_{k=1}^{F} \delta(q_k^{(z_i)}, q_k^{(z_j)}). \quad (3)$$

Two chunks with high perceptual similarity are prone to being confused in speech and likely to be dissimilar in terms of phonetic similarity. We select negative examples such that

$s_q \geq \frac{8}{11}$ using the approximate nearest neighbors approach as above. However, we pruned the selections to ensure that they were phonetically *dissimilar* with $s_{Ph} < \frac{8}{11}$ and thus not coincident with any positive examples.

4.3. Acoustic and Prosodic similarity

In addition to phonetic similarity, we also use low-level prosodic characteristics to identify similar pairs. In particular, we use per-frame acoustic intensity, fundamental frequency, and periodicity to evaluate the similarity between chunks, as extracted by the AuToBI toolkit [17]. AuToBI measures intensity in decibels, frequency in Hz, and periodicity as a number between 0 and 1. The frame rate of measurements is different between AuToBI and our system, so we linearly interpolate these values.

In order to obtain the intensity similarity between two chunks, we measure the Euclidean distance between their intensity vectors, I_i, I_j, and formulate the similarity s_I as:

$$s_I(I_i, I_j) = e^{-\alpha_I \|I_i - I_j\|}. \tag{4}$$

The exponential function is used so that the similarity between two chunks is one if the distance is zero and is zero if the distance between two chunks is very large. We use the same approach to derive the similarity for the fundamental frequency and periodicity with the parameters α_F and α_{Pe} respectively. In addition, we replace the frequency of voiceless or silent frames with 0.1, which is far from any real frequency, so that the similarity between two unvoiced frames is one. Finally, the overall similarity between two chunks, $s(z_i, z_j)$, is the weighted combination of similarities in phonetic content (s_{Ph}), intensity (s_I), fundamental frequency (s_F), and periodicity (s_{Pe}):

$$\begin{aligned} s(z_i, z_j) =& \lambda_{Ph} s_{Ph}(z_i, z_j) + \lambda_I s_I(z_i, z_j) \\ &+ \lambda_F s_F(z_i, z_j) + \lambda_{Pe} s_{Pe}(z_i, z_j). \end{aligned} \tag{5}$$

The weights $\lambda_{Ph}, \lambda_I, \lambda_F, \lambda_{Pe}$ lie in the interval $[0, 1]$ and are constrained to sum to one. The parameter values are found by tuning the system with a Bayesian optimization method described in Section 4.3.1.

4.3.1. Hyper-parameter search

We select the appropriate values for λ_{Ph}, λ_I, λ_F, λ_{Pe}, α_I, α_F, and α_{Pe} utilizing Spearmint [18], a Bayesian optimization package. It performs an intelligent hyperparameter search, attempting to minimize the frame-wise error rate shown in (6) on the development set. The parameters are searched within the following ranges: $\alpha_I \in [0.1, 0.3]$, $\alpha_{Pe} \in [8.0, 12.0]$, $\alpha_F \in [0.03, 0.07]$. These ranges were chosen base on the distributions of the distances measured for each feature type on the training data. In addition, the weights, λ_{Ph}, λ_I, λ_F, λ_{Pe} are also selected by the package in the range $[0, 1]$ and then divided by their sum.

4.4. Evaluation

We refer to the systems above as baseline (original system described by Mandel et al. [2]), phonetic (section 4.1), perceptual (section 4.2), and prosodic (section 4.3). We evaluate these systems using both objective (computed) and subjective (human-based) error metrics.

4.4.1. Objective computed metrics

Given a noisy speech signal x_j, the concatenative resynthesis system uses the learned similarity metric to identify a matching clean signal z_i from the dictionary. To evaluate the quality of the mapping, $x \rightarrow z$, we can compare the frame-level phonetic transcriptions of the x and z. We consider two objective metrics for evaluation: frame-wise error rate and phone error rate.

We compute the *frame-wise error rate* as:

$$e_f(z, x) = 1 - \frac{1}{F} \sum_{k=1}^{F} \delta(p_k^{(z)}, p_k^{(x)}), \tag{6}$$

where p_k is the phonetic label of the k^{th} frame in a signal. Thus, e_f considers the frame-wise phonetic correspondence of the input and output chunks. We used this metric when tuning our hyperparameters and building a final system for testing.

We also compute a *phone error rate* based on collapsing repeated phones, aligning the phonetic sequences of the input and output signals, then considering the insertion, substitution, and deletion errors. Since consecutive chunks overlap by 10 frames, we need to account for overlapping phone labels for frames during resynthesis. We do this by relabeling a frame with a composite phone when overlapping frames have different labels. For example, if two overlapping frames have different labels, say "B" and "D", we label the corresponding frame in the resynthesized chunk as "B-D" (the composite labels are always sorted lexicographically). After preprocessing the phone labels as described, we compute phone error rate by using the Speech Recognition Scoring Toolkit (SCTK) [19].

The frame-wise metric measures performance in terms of direct mappings between inputs and outputs of our system. The phone error rate collapses repeated sequences and takes alignment into account. Intuitively, we expect the latter metric to be higher and a rough approximation to an error rate if this system were used a preprocessor for transcribing words.

4.4.2. Subjective intelligibility metric

We also conducted listening tests determine a *subjective intelligibility error rate*. This metric measures how accurately listeners perceive what was spoken. We selected 12 random (noisy) utterances from the test set at random SNR levels. These were resynthesized to produce clean speech for each of the four systems (called baseline, phonetic, perceptual, and prosodic). We recruited three participants: two females and one male (ages 33, 34, and 35). Two were native speakers of English and one was a non-native, but fluent, English speaker. Participants were asked to listen to randomly selected examples of clean speech or resynthesized speech from one of the four systems, comprising a total of 72 stimuli. They were instructed to transcribe the spoken sentence for each utterance. Although they were provided with the GRID grammar, they were free to transcribe what they heard independently. The subjective error rate was measured as the number of words incorrectly transcribed.

5. Results

The phonetic and perceptual systems were tuned on the validation set using a grid search over the number of paired examples and a transition parameter (γ) [2] for mapping distances to affinities. The final phonetic system was trained using 800,000 positive and negative examples and $\gamma = 1.0$, while the final perceptual system was trained using 1,500,000 positive and negative examples and $\gamma = 1.0$. The prosodic system was tuned using the hyper-parameter search described in Section 4.3.1. The final tuned prosodic system was trained using the parameter values: $\lambda_{Ph} = 0.382$, $\lambda_I = 0.244$, $\lambda_{Pe} = 0.374$, $\lambda_F = 0.0$,

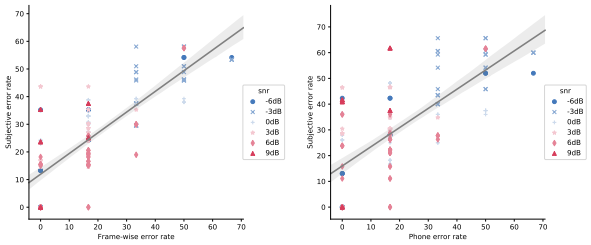

Figure 1: *Objective error rates (y-axis) versus subjective error rates (x-axis) from intelligibility tests at different SNR levels. Left: objective frame-wise error rates; right: objective phone error rates.*

Table 1: *Average objective frame-wise error rates for different systems and SNR levels.*

system	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg
prosodic	42.8	34.0	28.9	23.4	24.6	23.2	29.5
phonetic	44.3	34.7	30.5	25.3	27.8	24.2	31.1
perceptual	45.3	38.0	29.5	27.8	28.3	25.0	32.3
baseline	49.1	36.7	31.9	28.4	30.1	23.7	33.3

$\alpha_I = 0.3$, $\alpha_{Pe} = 8.0$, $\alpha_F = 0.03$; with 200,000 positive and negative examples. These parameter values indicate that the similarities due to the phonetic and periodicity characteristics of the signal contribute the most value, while the frequency information contributes the least. This may be since Autobi determined frequencies of both unvoiced consonants and silence as "NaN" and therefore the frequency was not as informative as other features.

The final results from resynthesizing the test set are shown in Table 1 (frame-wise error rates) and Table 2 (phone error rates). When compared to the baseline, our methods perform better better in almost all cases. While results based on our validation set had shown the perceptual system outperforming the phonetic system, the test set results in Tables 1 and 2 indicate that these systems have similar performance. The prosodic systems performs best based on the frame-wise metric, the metric used for tuning. As expected, the phone error rates are higher overall. In Table 2, using the phone error rate, we see that the perceptual and phonetic systems score better than prosodic. This may be because this metric was not used for tuning the systems. We also think of this metric as a rough approximation to a transcription phone error rate without a language model, and thus a loose upper bound on an objective error rate.

Intelligibility results from listening tests on random subsets of clean and resynthesized test utterances are summarized in Table 3. Clean speech utterances act as a control and naturally receive the lowest subjective error rates. For resynthesized utterances, participants generally achieved lowest error rates across different SNR levels for the prosodic system, which combines phonetic, acoustic and similarities. Participants achieved lower errors rates for the perceptual system relative to the baseline only at low and high SNR levels. This demonstrates that improving the training signal available to the DNN by combining acoustic, phonetic and prosodic characteristics of speech can yield noiseless signals that are more intelligible to humans.

To evaluate our computed metrics, we also looked at whether subjective error rates were correlated with the objective error rates. Figure 1 plots subjective errors against objective errors showing points at SNR levels. It also depicts a linear regression

Table 2: *Average objective phone error rates for different systems and SNR levels.*

system	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg
perceptual	60.8	48.6	36.2	32.8	31.0	28.4	39.6
phonetic	64.4	46.8	41.1	30.5	32.8	27.5	40.5
prosodic	71.6	56.2	46.8	39.4	39.2	41.3	49.1
baseline	95.4	78.2	65.6	64.4	66.3	59.4	71.6

Table 3: *Average subjective intelligibility error rates for different systems and SNR levels.*

system	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg
clean	0.0	0.0	0.0	0.0	3.7	0.0	0.6
prosodic	0.0	55.6	9.3	11.1	13.0	0.0	14.8
perceptual	16.7	44.4	24.1	11.1	20.4	0.0	19.4
phonetic	11.1	33.3	27.8	16.7	14.8	16.7	20.1
baseline	55.6	41.7	14.8	13.9	24.1	16.7	27.8

line of best fit to see how well intelligibility results may be predicted from computed error rates. We found an overall Pearson correlation of 0.78 between intelligibility and frame-wise error rates, and of 0.68 between intelligibility and phone error rates. The relatively high correlation demonstrates the promise of our computed error metrics, particularly frame-wise error rate, for automatic tuning and evaluation of a concatenative resynthesis system.

6. Conclusions

We have shown that concatenative resynthesis systems can make more efficient use of the available training data and maximize matching performance for the goal of producing noiseless and high quality speech from noisy signals. This is made possible through improved training signals for the similarity metric learning DNN by selecting paired examples that incorporate acoustic, phonetic, and prosodic characteristics of speech. In particular, we found that combining similarity based on phonetic content and periodicity of the signals yields best results. In our experiments, we show that 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. We also demonstrated the effectiveness of using computed error metrics for tuning and evaluating such systems. Moreover, we confirmed the correlation of computed error metrics with subjective human evaluations by conducting intelligibility listening tests. However, our system was developed and tested on small vocabulary and speaker dependent data. In future work, we will investigate the performance and scalability of our approach on large vocabulary [20] and speaker independent speech.

7. Acknowledgements

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.

8. References

[1] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on audio, speech,*

and language processing, vol. 14, no. 4, pp. 1218–1234, 2006.

[2] 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.

[3] 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.

[4] S. Maiti and M. I. Mandel, "Concatenative resynthesis using twin networks," *Proc. Interspeech 2017*, pp. 3647–3651, 2017.

[5] X. Xiao and R. M. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1243–1257, 2010.

[6] R. M. Nickel, R. F. Astudillo, D. Kolossa, and R. Martin, "Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 983–997, 2013.

[7] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822–836, 2011.

[8] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba *et al.*, "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 *Machine Listening in Multisource Environments*, 2011.

[9] A. Ogawa, K. Kinoshita, T. Hori, T. Nakatani, and A. Nakamura, "Fast segment search for corpus-based speech enhancement based on speech recognition technology," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1557–1561.

[10] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second chimespeech separation and recognition challenge: Datasets, tasks and baselines," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 126–130.

[11] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[14] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: trainable text-speech alignment using kaldi," 2017.

[15] L. Boytsov and B. Naidan, "Engineering efficient and effective non-metric space library," in *Similarity Search and Applications - 6th International Conference, SISAP 2013*, 2013, pp. 280–293.

[16] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[17] A. Rosenberg, "Autobi-a tool for automatic tobi annotation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[18] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[19] J. Fiscus, "Speech recognition scoring toolkit (sctk)." [Online]. Available: https://www.nist.gov/itl/iad/mig/tools

[20] S. Maiti, J. Ching, and M. I. Mandel, "Large vocabulary concatenative resynthesis," in *Proceedings of Interspeech*, 2018, to appear.