Language and Noise Transfer in Speech Enhancement Generative Adversarial Network

Santiago Pascual[1], Maruchan Park[2]
Joan Serrà[3], Antonio Bonafonte[1], Kang-Hun Ahn[2]

April 20, 2018

[1] Universitat Politècnica de Catalunya, Barcelona, Spain
[2] Chungnam National University, Daejeon, Republic of Korea
[3] Telefónica Research, Barcelona, Spain

Table of contents

Introduction

Speech Enhancement

- Speech enhancement tries to improve the intelligibility and quality of speech signals that have been distorted or contaminated by noise.
- It is deployed in a myriad of speech processing applications, specially as a pre-processing stage.
- It is important to work in a broad range of conditions to improve intelligibility and/or naturalness.
  - New languages.
  - New noises with different behavior.

Transfer Learning

- Transfer learning encompasses a set of techniques to adapt or fine-tune already pre-trained models (e.g., a first language) to a different but related task/domain (e.g., another language).
- We normally do transfer learning when we have less amount of data for the final task than in the pre-training.
- Transferring takes advantage of common features between domains (e.g., low level features extracted from waveforms).

Motivation

- We want to quantify how a speech enhancement GAN (SEGAN) performs on languages and noises for which it was not trained.
- We want to assess the amount of new data required to transfer what it has learnt to the new setting.

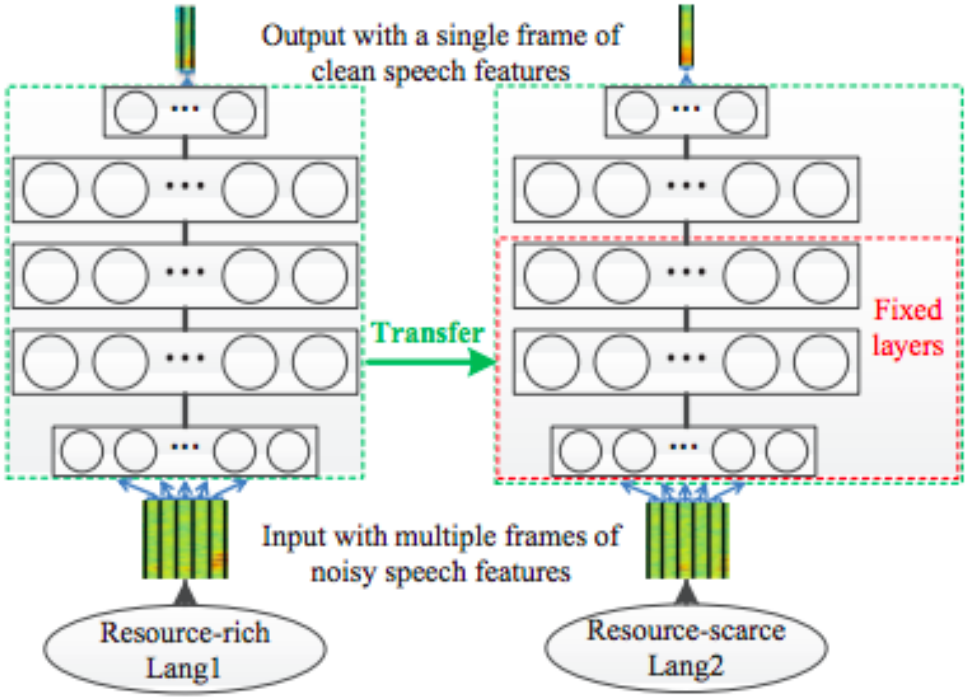
Motivation

- Does the system perform well for a language it has not been trained for?
- And for noises?
- If not, which is the amount of data necessary to adapt the system to the new language/noise?
- In both cases, how critical is the selection of training data for adapting the system to the new task?
- Is it worth to retrain from scratch or is it better to reuse a pre-trained GAN?

Related Work

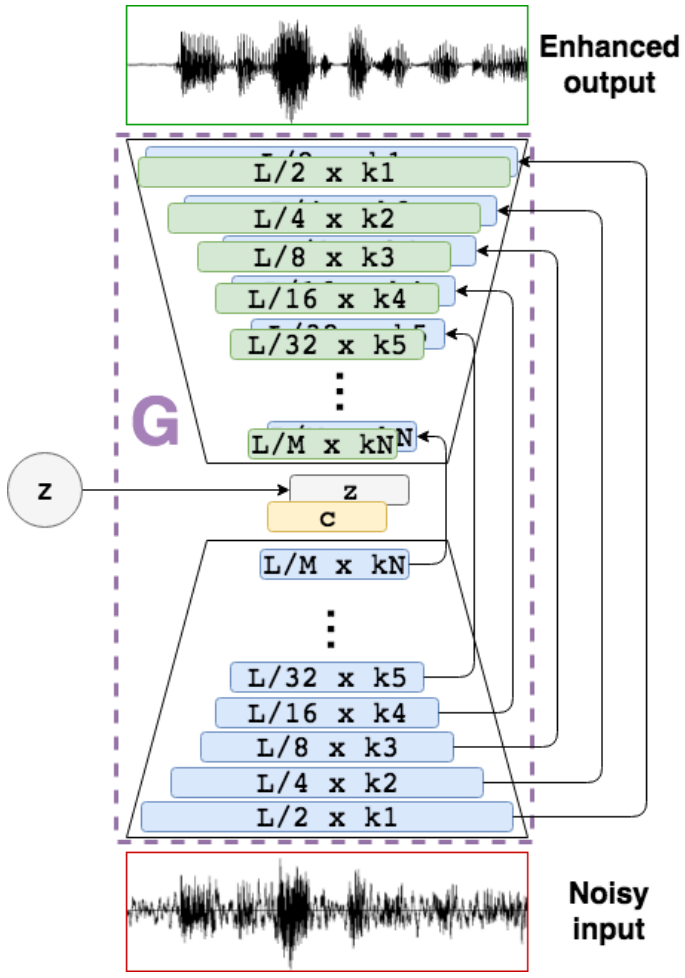
Related Work

Transfer learning applied to top-most abstraction layers from English to Mandarin in deep feed-forward neural networks as in Xu et al. 2014, where they needed 1 minute of adaptation speech to achieve good performance.



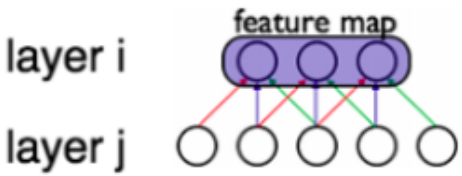
SEGAN Architecture

- Encoder (conv): Project noisy signal into the latent space.
- Decoder (deconv): Interpolate the intermediate hidden features with learnable params. until re-generation of clean speech.



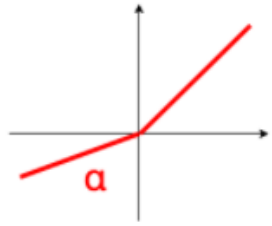
SEGAN Architecture

- 1D convolutional layers with PReLU activations (G) or LeakyReLU activations (D).



$$h_i^k = \tanh(W_{ij}^k * h_j + b_i^k)$$

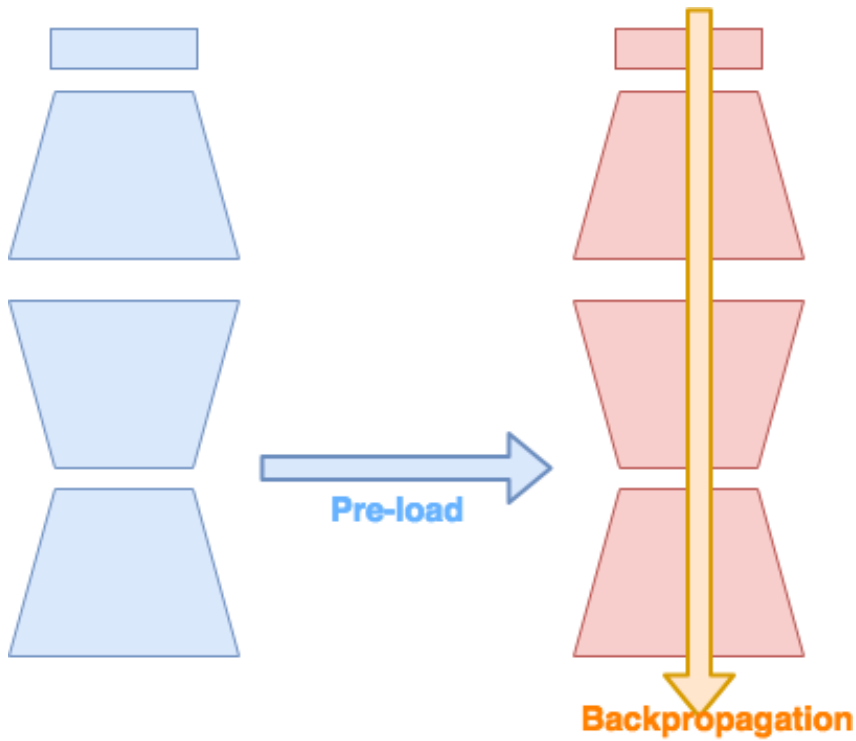
$$x$$



- Virtual batch normalization (Salimans et al. 2016) in D for stability and better backprop.
- Skip connections in G.
  Training is performed with the same schedule or true/fake data-pairs as in the original SEGAN work.

Transfer learning approach

To perform transfer, all G and D weights are loaded from a prior language to the new one, and every weight is retrained (we did not perform any layer freezing test in this work).



Experimental Setup

Objectives

- We investigate transfer learning from English to:
  - Catalan
  - Korean
- Experiment 1 trains SEGAN over different speech durations for two baselines: (1) pre-trained English (preeng) and (2) from scratch.
- Experiment 2 checks generalization to unseen test noises and variation of performance with training noise types.

Experiment 1

- Based on English and random initialization, SEGAN is trained over a range of speech durations: 24 s, 1, 2, 4, 10, 20, 50, 100, and 200 min.
- Models are trained with different seeds for:
  - 10 times in short durations: 24s, 1, and 2 minutes.
  - 5 times for larger durations.
- This way we observe performance differences between two init schemes with varying speech durations to explain transfer data requirements.

Experiment 2

- We focus on 20 min of data, as Experiment 1 showed it was enough to continue.
- Verification of performance with unseen noises is conducted training over different noise types.
- Train 5 times with each randomly selected noise, increasing amount of noises one by one until 10 (50 times in total).

English dataset

- 30 speakers (15 male, 15 female), with each speaker recording 400 sentences from the Voice Bank Corpus.
- 28 speakers used for training, mixed with 40 noise conditions (10 noise types, 4 SNRs: 15, 10, 5, and 0 dB).
- 2 speakers used to test with 20 noise conditions (5 noise types, 4 SNRs: 17.5, 12.5, 7.5, and 2.5 dB).

Korean dataset

- Recording 20 minutes per speaker for 12 speakers (6 male, 6 female) in a quiet room.
- Sentences chosen from the Korean web portal NAVER Open Podium and NAVER Encyclopedia.

Catalan dataset

- Recording 20 minutes per speaker for 12 speakers (6 male, 6 female). The recordings took place in a recording studio.
- Each speaker recorded at least 1h of short paragraphs (though just 20 min are taken to match Korean), which were selected from a set of novels to achieve phonetic and prosodic coverage.

Both datasets were contaminated with the same approach as in English, and we take 10 speakers to train and 2 to test.

Hyperparameters

- The English model is trained for 86 epochs, and all models are trained with the RMSprop optimizer with the same hyper-parameters as in the original SEGAN work.
- All Catalan and Korean models are trained with batches of 100 samples for 30 epochs.
- The amount of epochs was reduced to a third of the original setting for faster experimentation given the large amount of models.

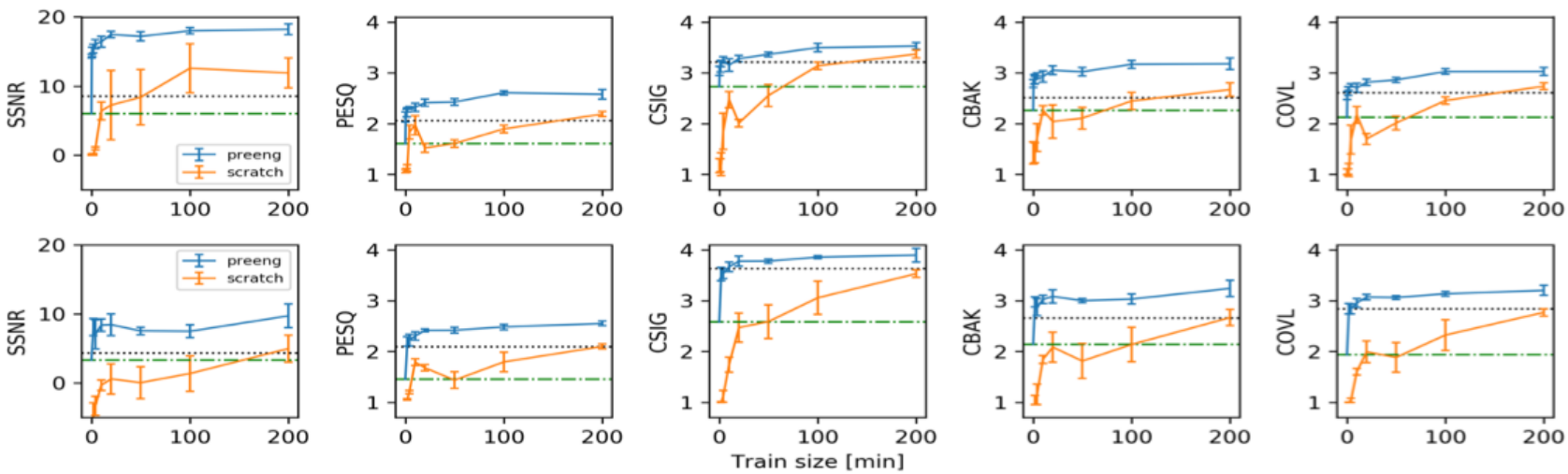
Evaluation metrics

- Perceptual evaluation of speech quality (PESQ): [-0.5, 4.5]
- Mean opinion score of signal distortion (CSIG): [1, 5]
- Mean opinion score of intrusiveness of noise (CBAK): [1, 5]
- Mean opinion score of overall effect (COVL): [1, 5]
- Segmental SNR (SSNR): [0, inf)

Results

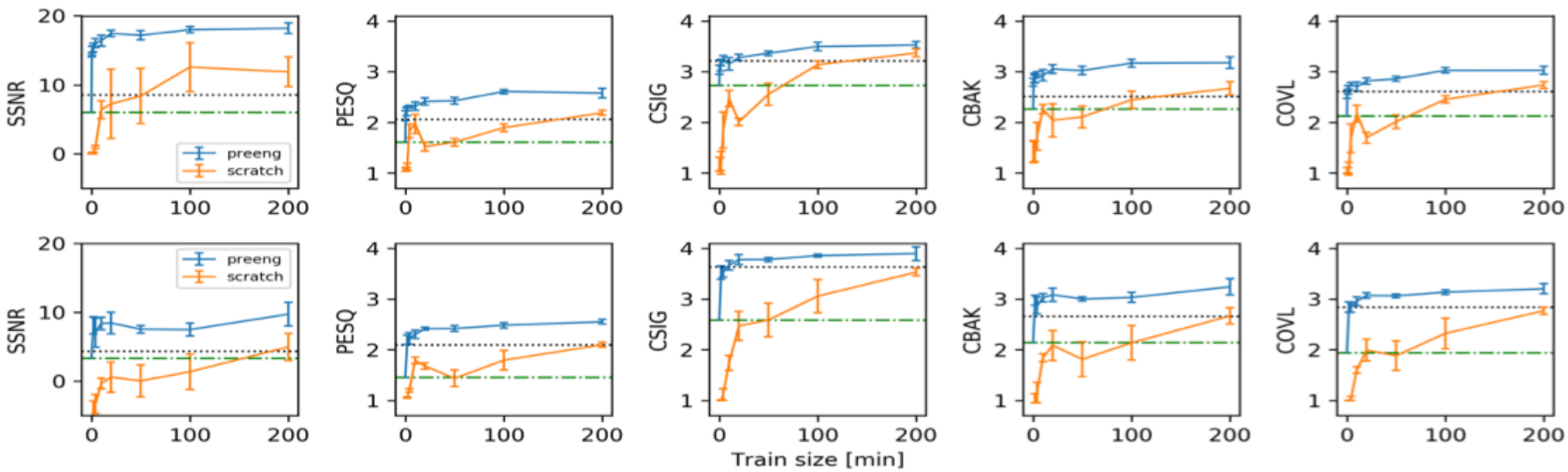
Language transfer results

- Catalan: top row.
- Korean: bottom row.
- Blue line: pre-trained English performance (preeng).
- Orange line: trained from scratch.
- Green dashed line: SEGAN without tunning.
- Black dash-dotted line: noisy level.



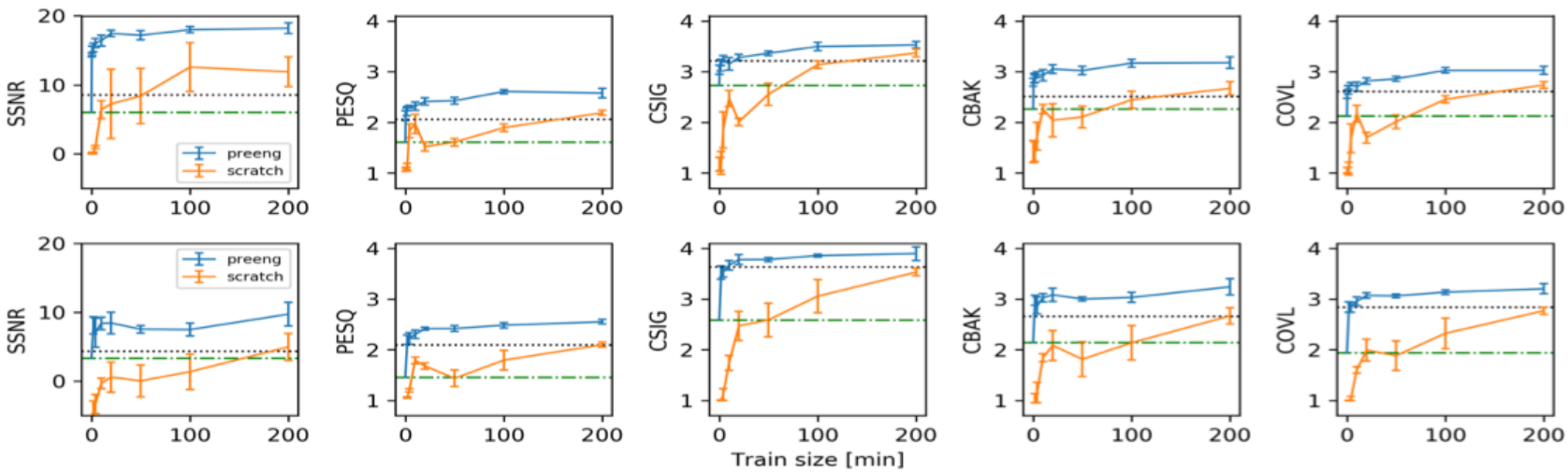
Language transfer results

- Despite the important differences between Catalan and Korean, the results for the two languages show very similar trends.
- Pre-trained English system alone does not perform well with the new languages.



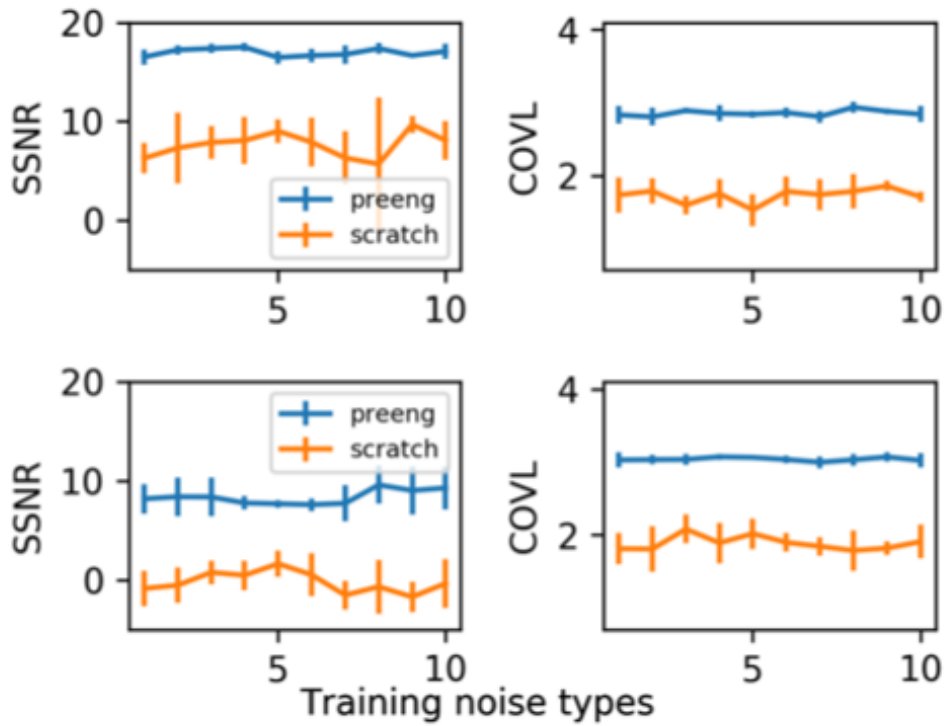
Language transfer results

- More importantly, only few minutes of new training data drastically improve performance. Even the least amount considered (24 s) significantly improves test performance.
- All metrics show a knee at ≈10 min, a threshold from which having more data shows diminishing returns.
  - Also amazingly small compared to English training time (9.4 h).



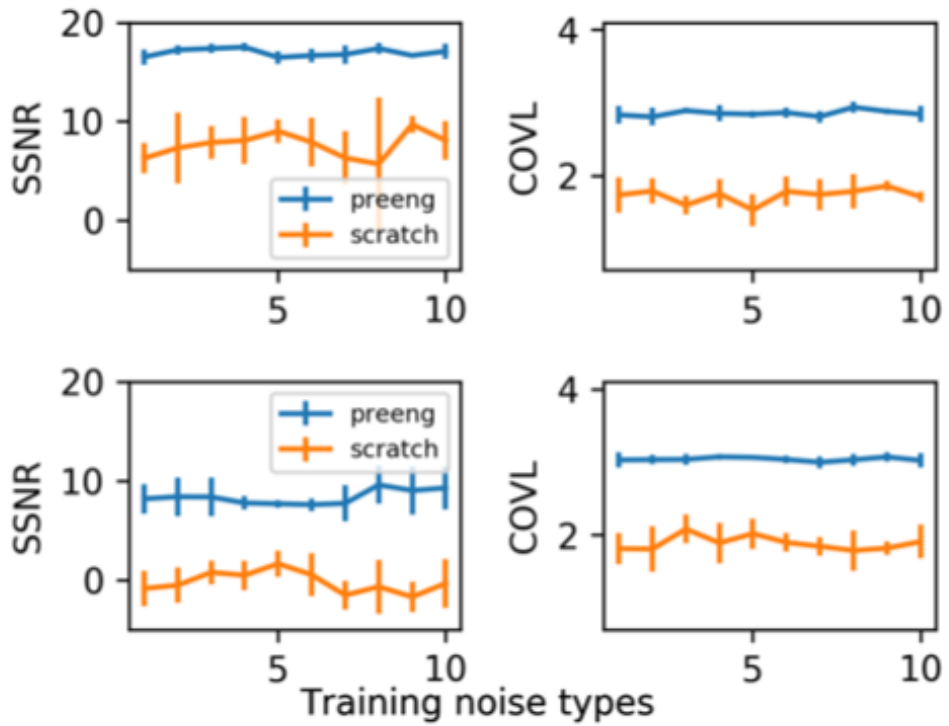
Noise experiment results

- Catalan: top row.
- Korean: bottom row.
- Blue line: pre-trained English performance (preeng).
- Orange line: trained from scratch.
- Qualitatively similar plots were obtained for PESQ, CSIG and CBAK.



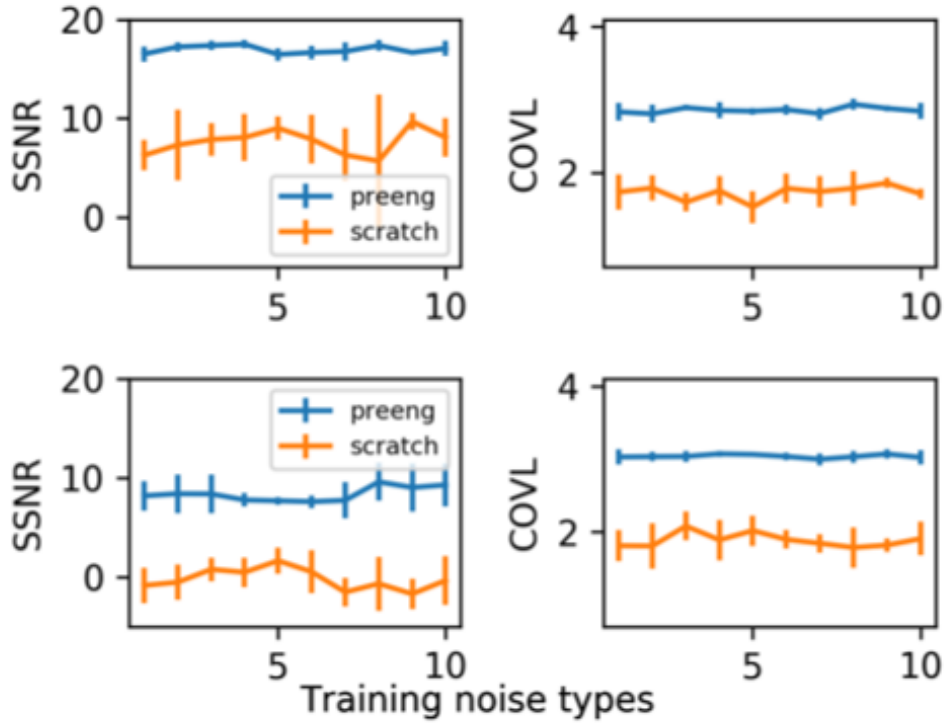
Noise experiment results

- Objective test metrics do not present a dependence on the number of types of training noise.
- Whether the training is with English pre-training or from scratch, performance to unseen noises is not affected by the amount of training noises.



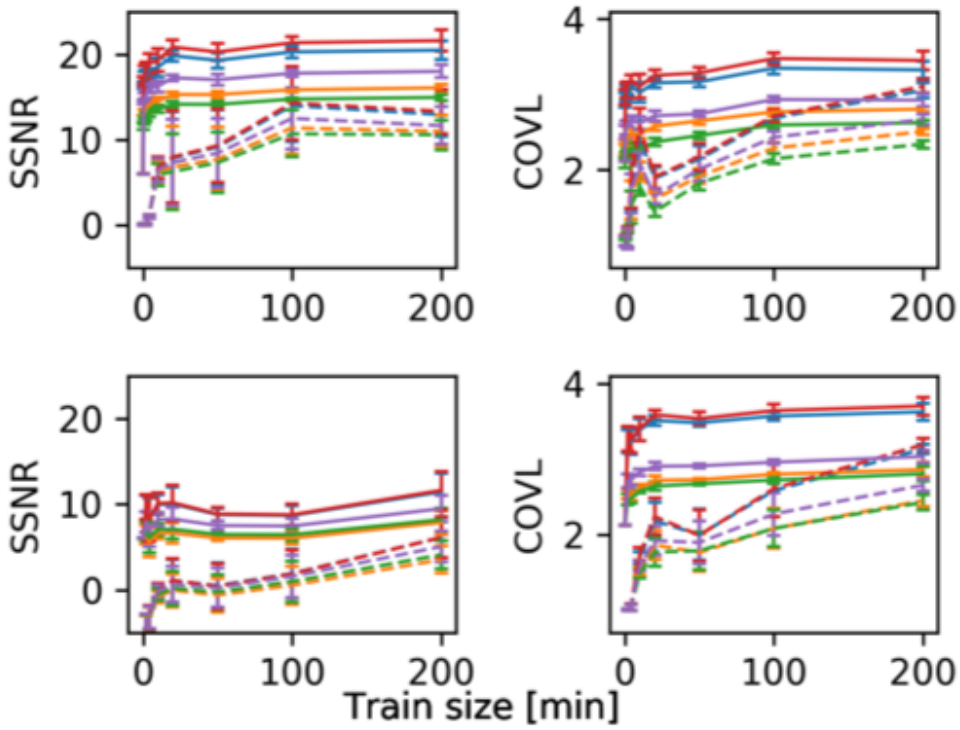
Noise experiment results

- Only difference b/w both versions seems to be in the variances, with the scratch version presenting a much larger variance.
  - In the case of training from scratch, one should be careful in which types of noise are considered for training.



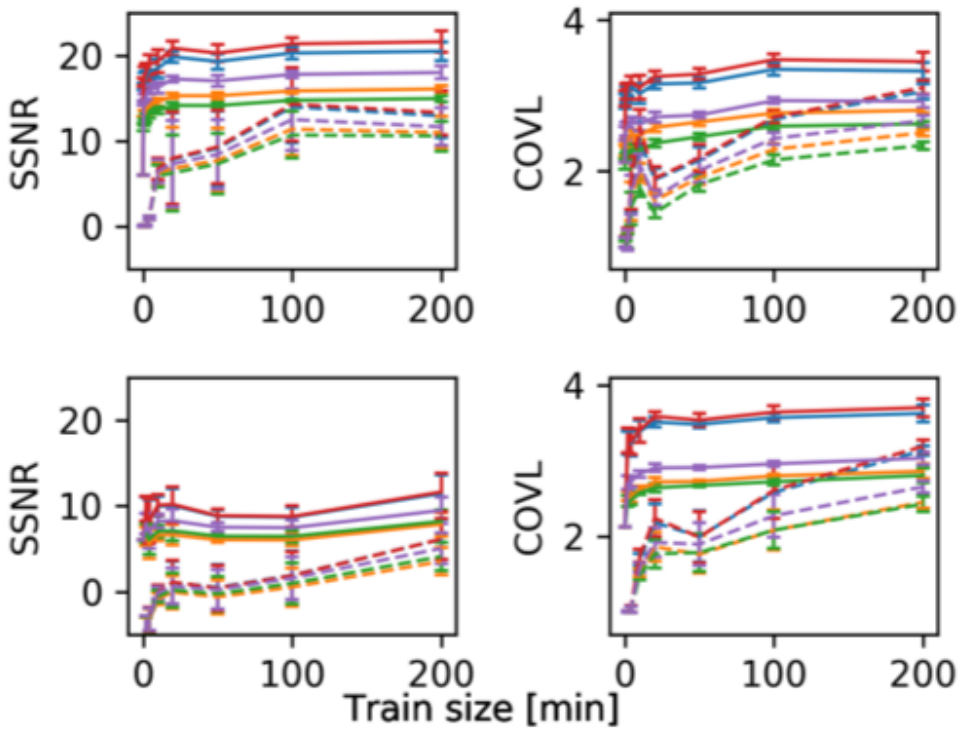
Results on different test noise types

- From top to bottom (first solid, then dashed lines): office (red), bus (blue), street (purple), living room (orange), and cafe (green) noises. Solid are preeng, dashed are scratch.
- Qualitatively similar plots were obtained for the PESQ, CSIG, and CBAK metrics.



Results on different test noise types

- Also observe consistent behaviors between languages and metrics.
- Office and bus noise types performing best and street noise, living room, and cafe noises (in this order) performing worse.
- Office and bus noise seem to ‘cluster’ in some metrics (e.g., COVL) in the upper side of the plots, while street noise, living room, and cafe noises also ’cluster’ in lower metrics.



Conclusions

Conclusions

- Transfer learning is very efficient for inter-language speech enhancement by a generative adversarial network.
- Pre-trained SEGAN with English achieves high performance even for short training time of Catalan and Korean (24 s), with unseen speakers and noise.
    - Adaptability to low resource environments.
- The number of noise-type in training is not crucial factor for the performance of the speech enhancement GAN.

Thanks!