Raytheon
BBN Technologies

Two Stage Data Augmentation for Low-Resourced Speech Recognition

W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R.Schwartz

Raytheon BBN Technologies

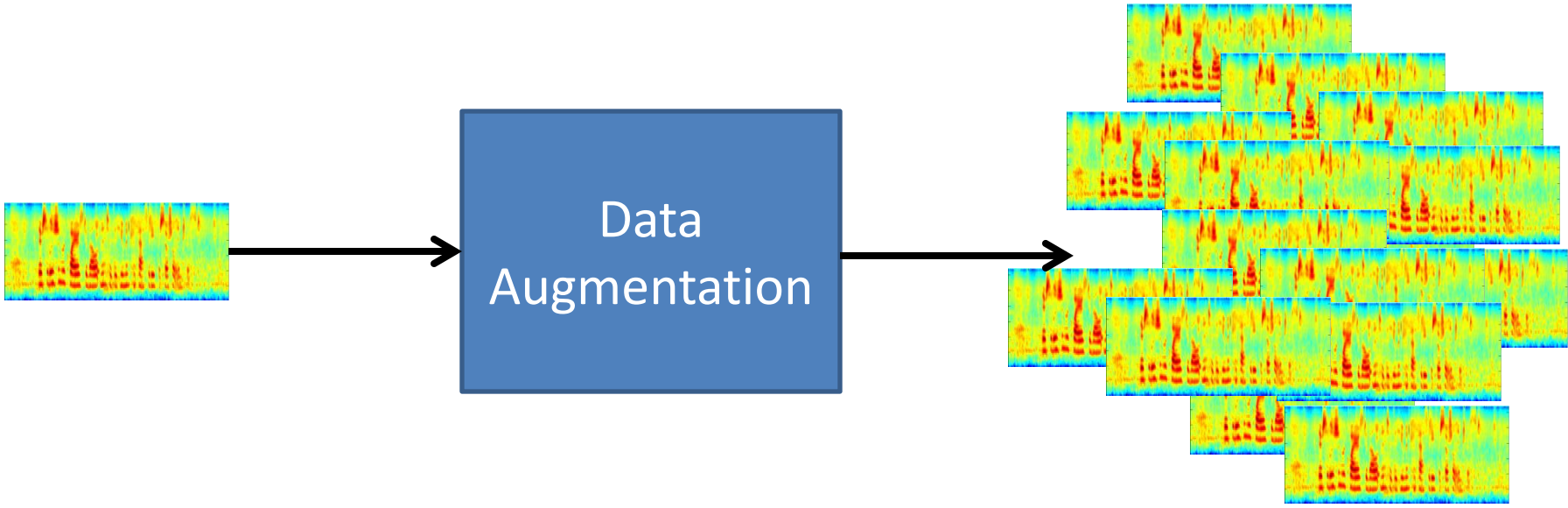
This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

Motivation

- For many languages, we have a limited amount of audio data.
- The problem is not just the limited total, but also the limited types.
  - Small number of speakers.
  - Under-represented dialects.
  - Limited environmental and recording conditions.

Goal

- Given a set of audio data, apply some transformations to produce more audio.



Methods

- Additive Noise
  - Add noise from non-speech segments of Babel data
- Speed Perturbation [Ko et al., 2015]
  - Stretch or contract the audio by a small factor, altering speaking rate, pitch, and formant frequencies
- Reverberation
  - Simulate audio being recorded in a reverberant environment
- Speaker-based Transformation [Cui et al., 2014]
  - Linear transformation to give speech characteristics of another speaker

IARPA Babel Data

- Conversational telephone speech collected in a variety of environments.
- 40 hours of transcribed audio.
  - Used to both LM and AM training.
- Four Languages from the fourth year.
  - Amharic, Guarani, Igbo, Pashto
- Results reported on a 10-hour development set.

Experimental Setup

- We use the new BBN Sage speech processing toolkit.
  - Incorporates elements from Byblos, Kaldi, CNTK, and other toolkits.
  - Poster at 4pm today (New Products and Services)
- DNN acoustic model with 6 hidden layers.
- fMLLR-transformed bottleneck features.
- Lexicon derived from simple G2P rules.

First Stage Data Augmentation Settings

- Noise Augmentation
  - Selected silence regions from previous Babel languages.
  - Mixed with the original data at a random SNR between 0 and 20 dB.
- Speed perturbation
  - Applied a random speed factor between 0.9 and 1.1 to each copy.

Preliminary Results

- Comparison of augmentation techniques

Language	Augmentation Type x Copies	WER
Amharic	None	44.2
Amharic	Speed x 2	44.0
Amharic	**Noise x 2**	**43.4**
Amharic	Reverb x 2	43.8

- Combining augmentation techniques

Language	Augmentation Type x Copies	WER
Amharic	Speed x 1, Noise x 1	43.5
Amharic	**(Speed+Noise) x 2**	**42.8**
Amharic	(Speed+Noise+Reverb) x 2	43.4

First Stage Results across Languages

Language	Baseline	One Copy	Two Copies
Amharic	44.2	43.4	42.8
Guarani	46.7	45.6	45.2
Igbo	55.5	54.5	54.3
Pashto	48.1	46.8	47.1
Average	**48.6**	**47.6**	**47.4**

- For all four languages, the data augmentation improves performance
- An additional second copy provides small gains for all languages except Pashto
  - A possible reason is the lack of microphone data for Pashto
- Based on these results, we used two copies of speed and noise augmented data for the dev language evaluations

Stochastic Feature Mapping (SFM)

- Cui et al., 2014 proposed SFM for Babel.
  - Two linear transforms map one speaker's features to another speaker's space.
- Set of speakers S.
- Set of speaker-dependent transforms T.
  - $T_{i,j}$ transforms S_i to S_j feature space.
- Set of fMLLR transforms F.
  - F_i transforms S_i to canonical feature space.
- SFM: $S_i * T_{i,j} * F_j = S'_i$

fMLLR-based Augmentation (FBA)

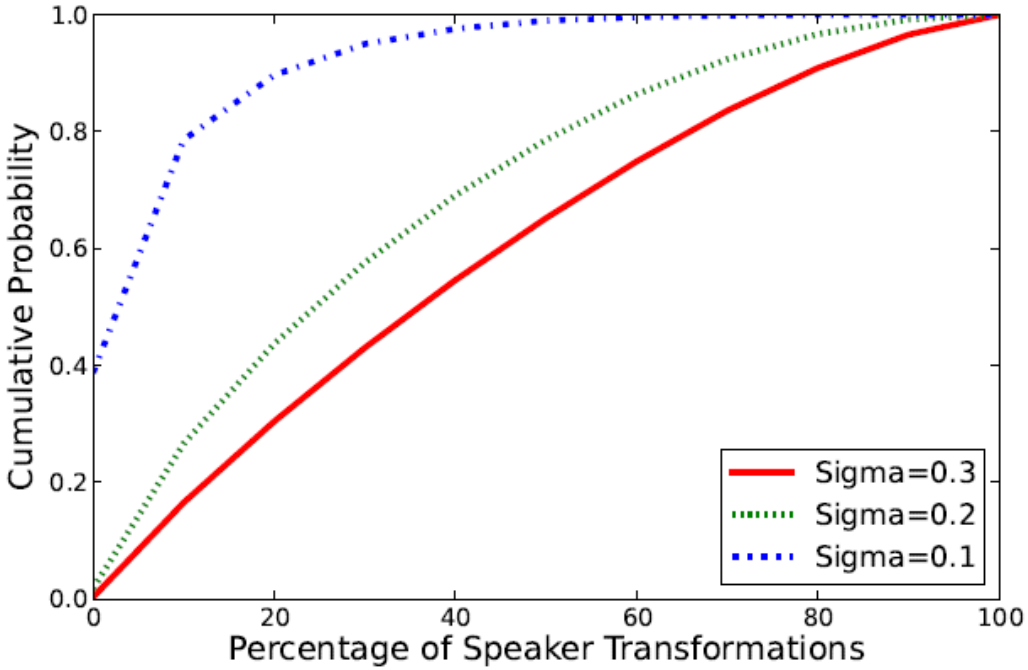
- Simplify by removing the speaker-dependent transform.
- Motivation is that just like noise or speed augmentation, we are just perturbing the features.
- Set of speakers S
- Set of fMLLR transforms F
- FBA: $S_i * F_j = S'_i$

Introducing a Control Parameter

- Noise and speed have control parameters (SNR, speed factor).
- We want a similar control parameter.
- Compute the similarity between two speaker transforms.
- Let the probability of selection be proportional to the similarity.
- Can control the level of perturbation by σ.

$$sim(A, B) = \exp\left(\frac{-\|A - B\|^2}{2\sigma^2}\right)$$

Introducing a Control Parameter



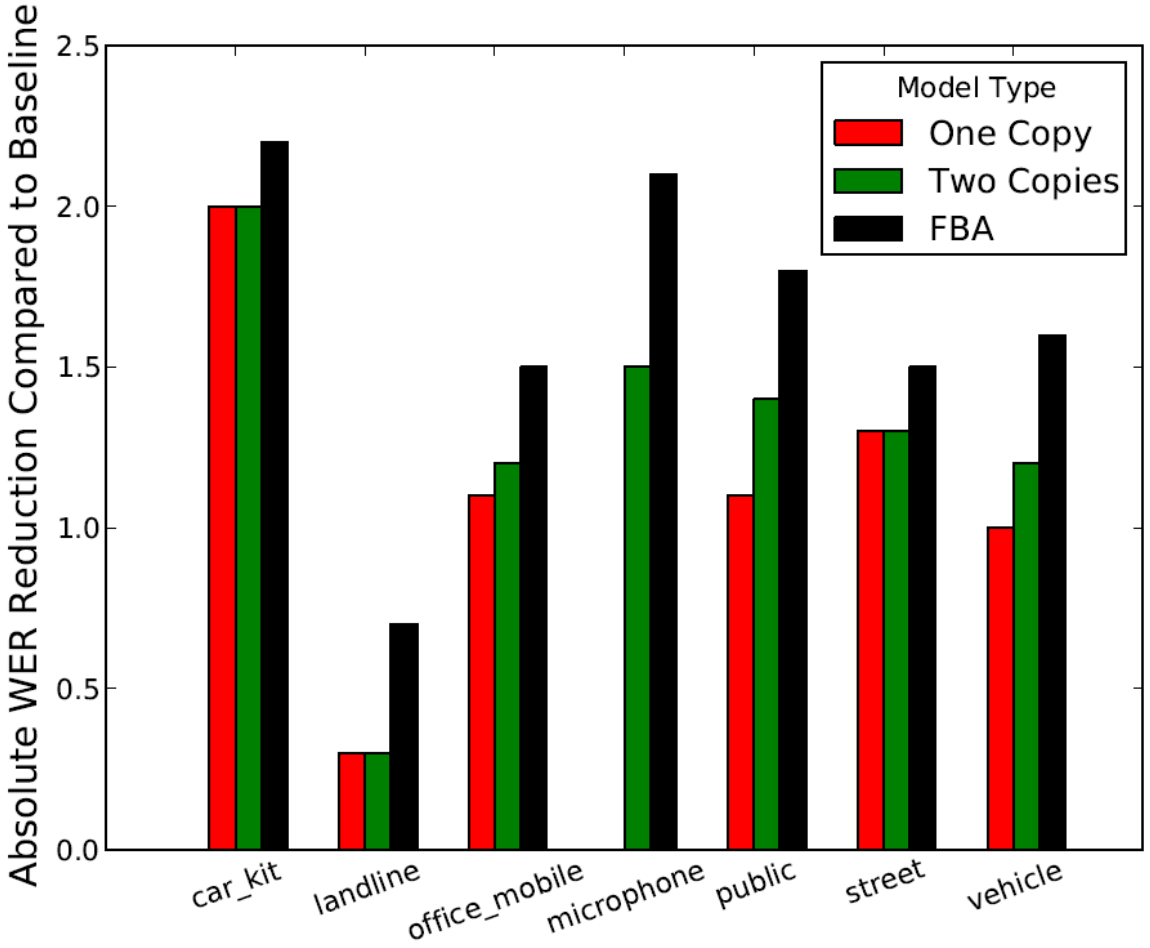
- Small values of σ bias the augmentation to more similar speakers.
- Large values of σ give a more uniform distribution.

Second Stage Results

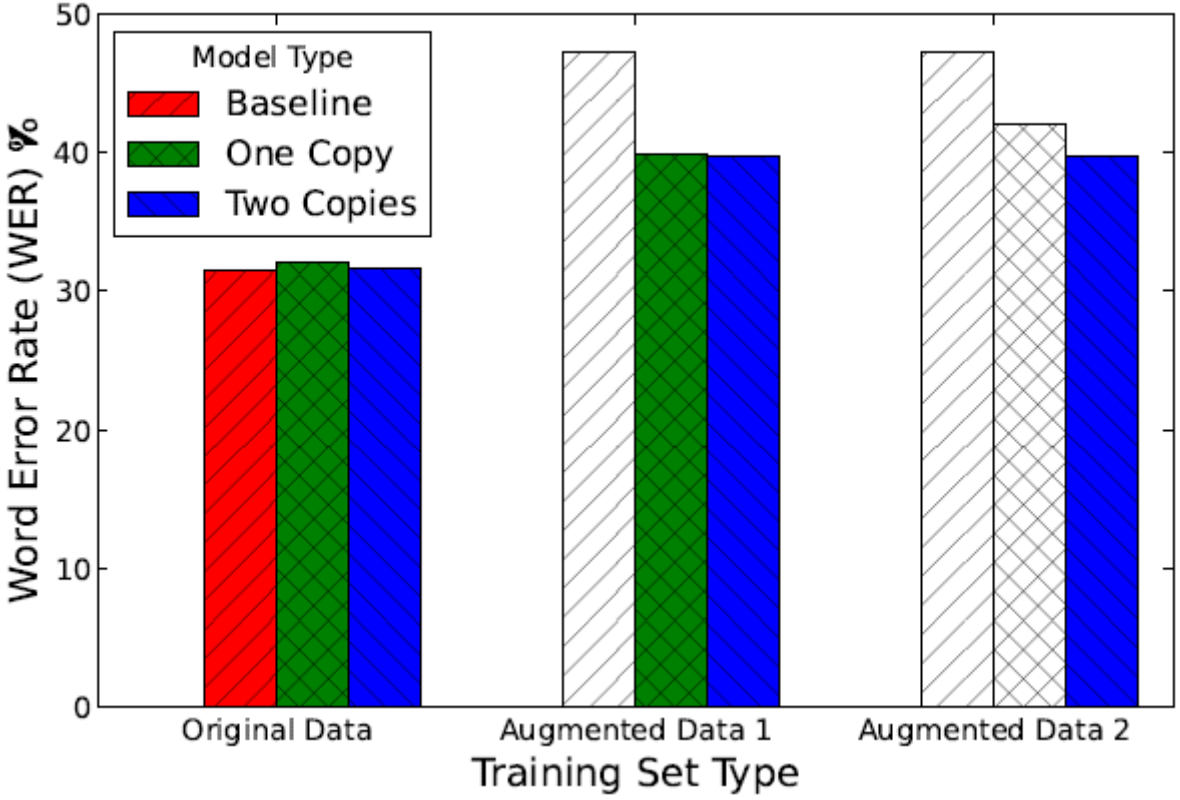
Language	First Stage	SFM	FBA; Random	FBA; σ=0.2
Amharic	42.8	42.6	42.4	42.2
Guarani	45.2	44.7	44.9	44.6
Igbo	54.3	54.0	54.1	53.9
Pashto	47.1	46.7	46.8	46.7
Average	**47.4**	**47.0**	**47.1**	**46.9**

- The second stage augmentation is applied only to the augmented copies from the first stage.
- All approaches give a similar small gain over the first stage.

Performance across Recording Conditions

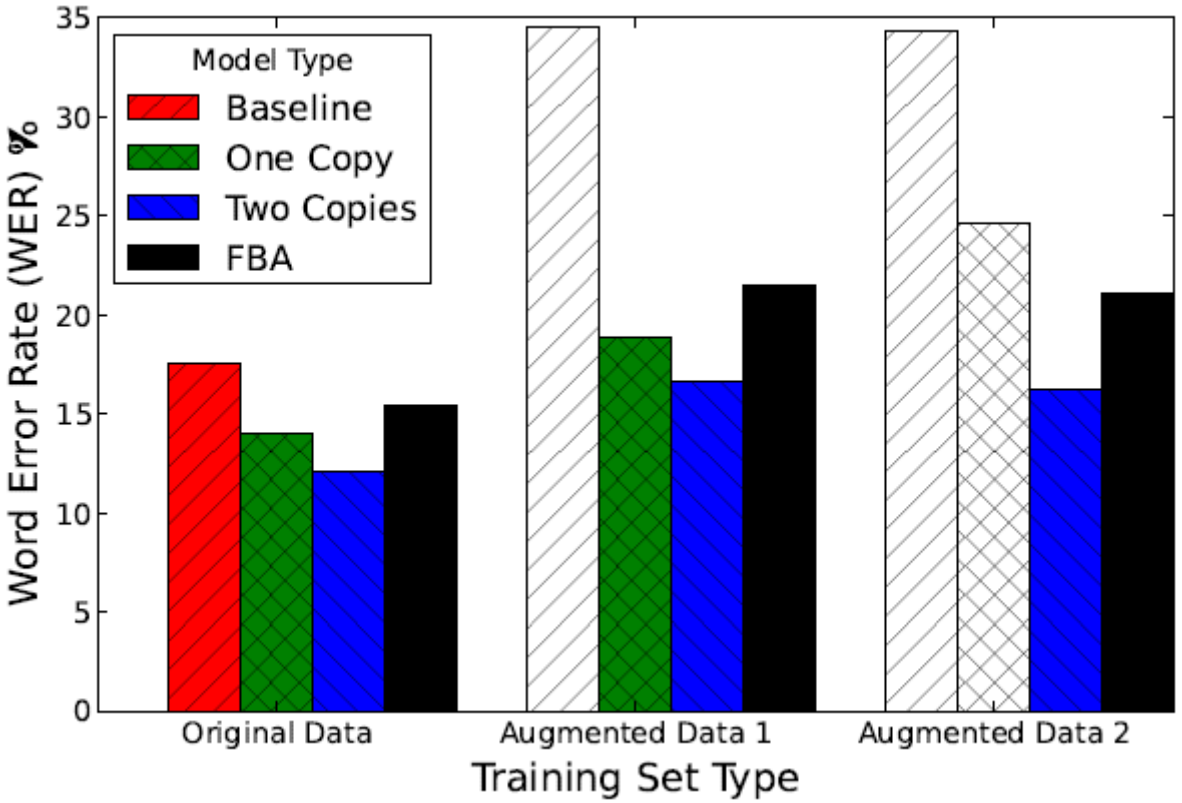


Performance Analysis for GMM on Training Data



Data augmentation does not help GMM on clean data.

Performance Analysis for DNN on Training Data



- Data augmentation improves performance on original data.
- FBA decreases performance on train, but improves on test.

Conclusions

- Introduced a second stage augmentation (FBA) where the level of perturbation can be controlled by a single parameter.
- Used first stage during Babel evaluation.
  - Improved best monolingual result (joint decoding of DNN+CNN+BLSTM) by 1.5% WER and 2.0% ATWV.
  - Also, improved performance with multilingual features by at least 1% WER.