Adaptation and Frontend Features to Improve Naturalness in Found-Data Synthesis

Erica Cooper and Julia Hirschberg

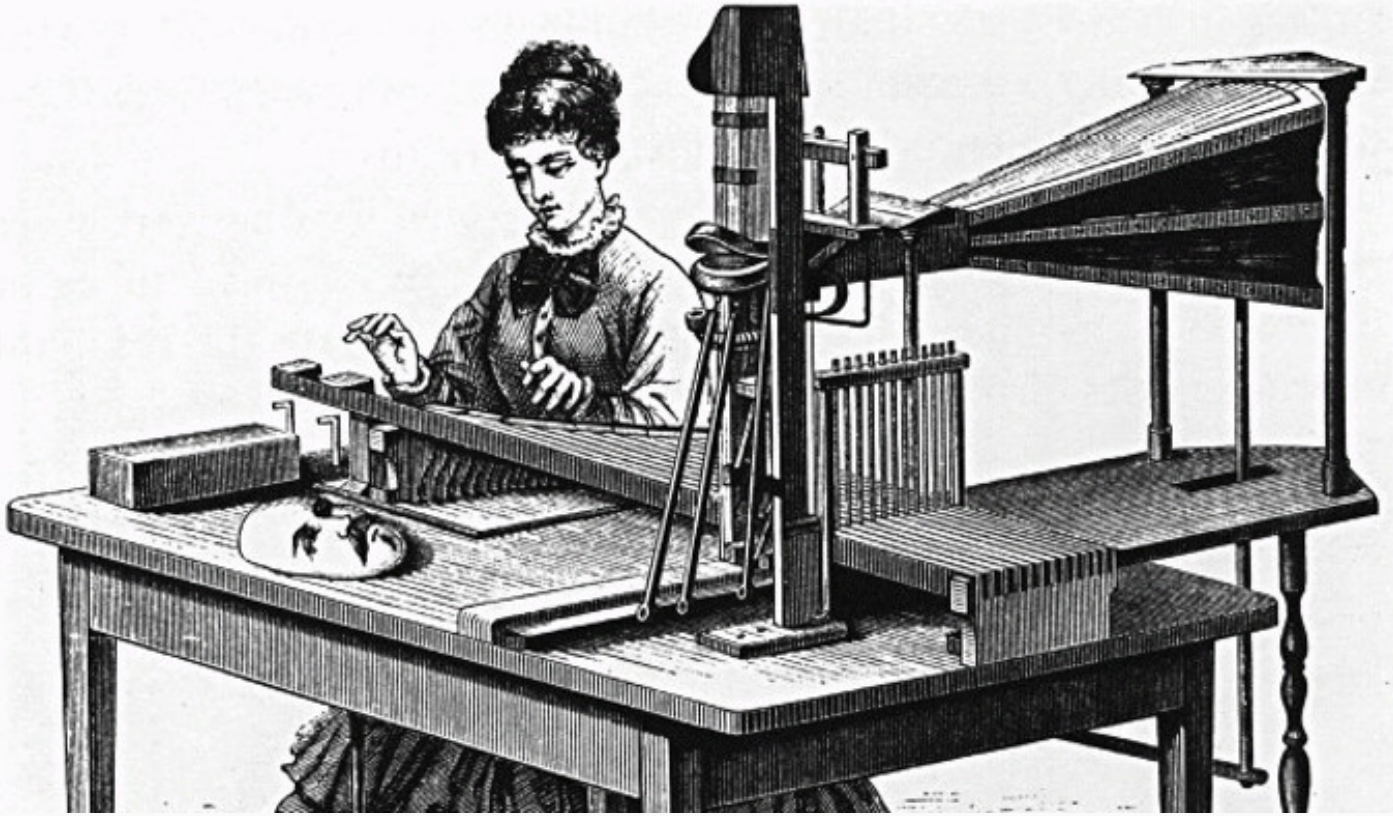
June 16, 2018

Spoken Language Processing Group
Department of Computer Science, Columbia University

Table of contents

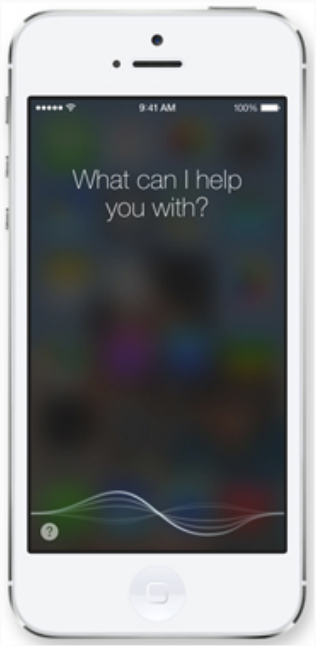
Introduction

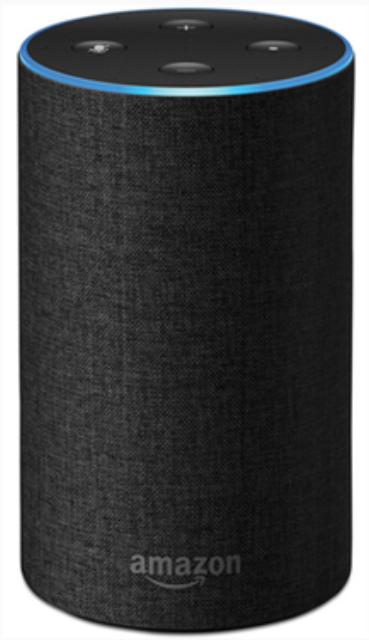
Speech Synthesis



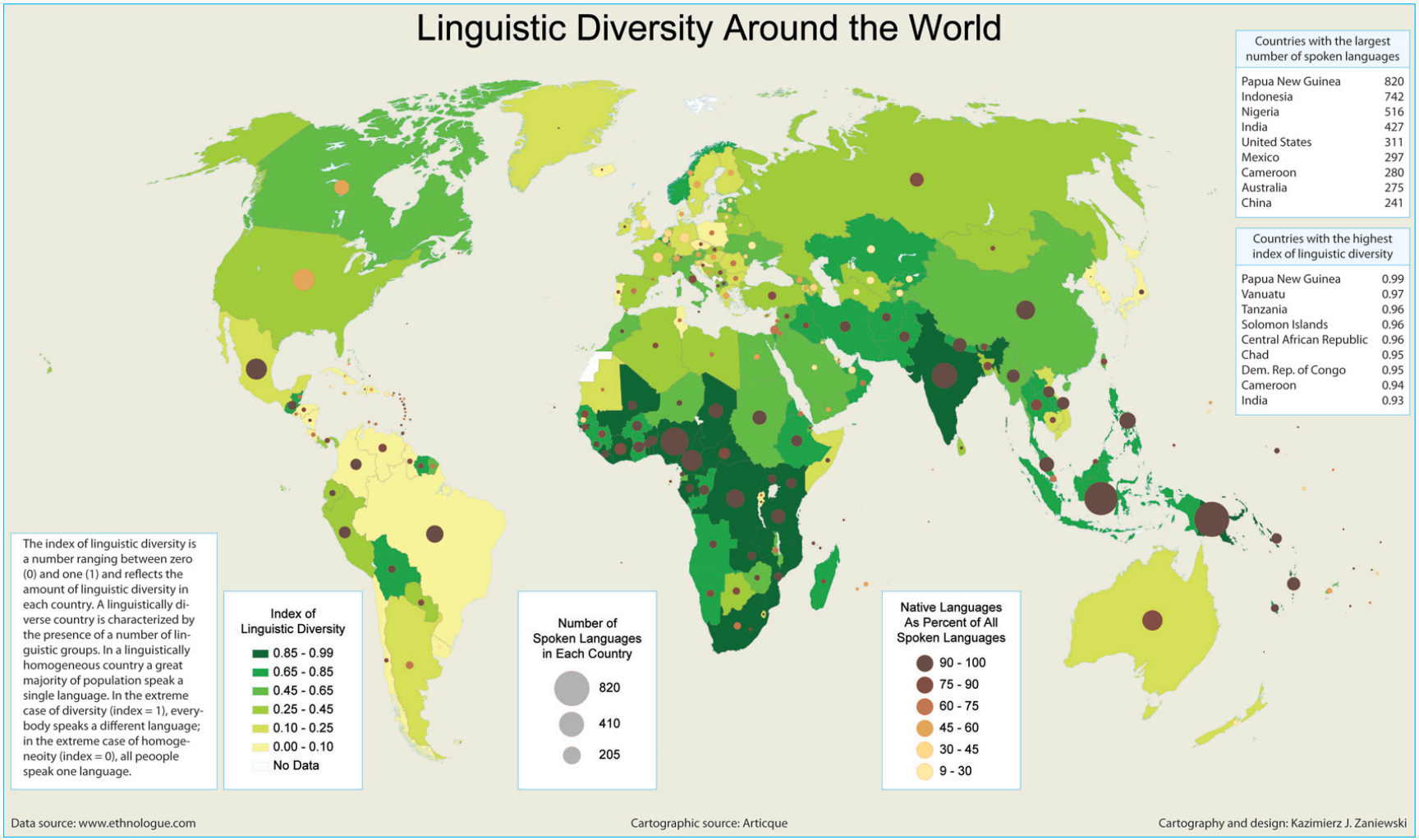
Joseph Faber's "Euphonia" (1846)

Speech Synthesis



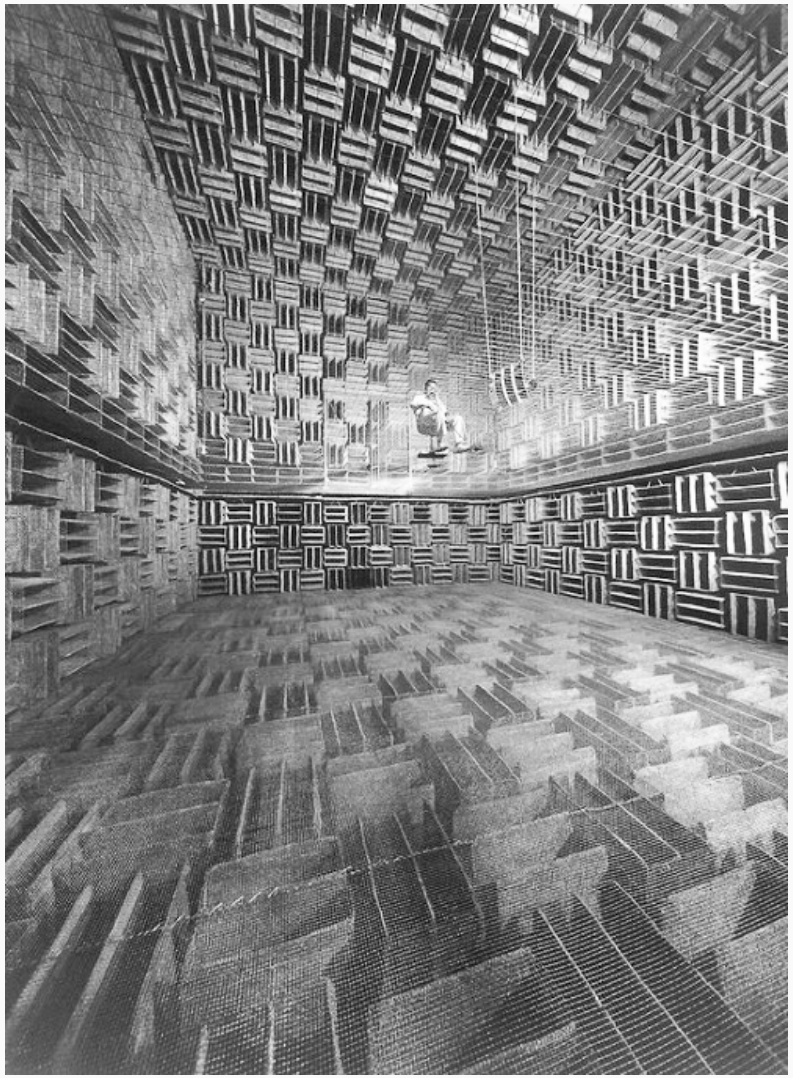


Linguistic Diversity



Source: https://journals.openedition.org/confins/6529

TTS Data



Bell Labs Anechoic Chamber, 1947 (Photo credit: Eric Schaal)

Found Data

Project Overview

Goal: To create TTS voices using high-quality found data and to optimize them for naturalness.

Approach: Include knowledge of additional acoustic and prosodic features in the voice modeling process in order to make voices more similar to the neutral TTS style, without sacrificing any of the data.

- HMM-based and neural network based synthesis
- Adapt models towards different subsets of the data selected based on acoustic / prosodic features
- Include new feature information at the frontend to enable selection of speaking style at synthesis time

Approaches and Related Work

Tools and Data

Tools: HTS (HMM synthesis), Merlin (NN synthesis), Festival (frontend), Praat and SoX (feature extraction)

Data: BURNC. High quality broadcast news speech. 3 female speakers; 4hrs 22min data.

Features: Utterance-level mean and sdev of f0 and energy; speaking rate; level of articulation; utterance length.

Subsets: Utterances are marked as having a high, medium, or low value for each feature, where each partition comprises a third of the total data.

Training Data Subset Selection

Popular approaches include (Stan et al. 2013, Chalamandaris et al. 2014, Cooper et al. 2016):

- Discard utterances with low alignment confidence
- Select for neutral-style utterances
- Select for utterances with similar recording conditions
- Remove utterances that are outliers

This approach makes sense when there's a larger amount of noisy or mixed-style data.

For smaller amounts of high-quality, relatively neutral data like BURNC, we would like to see if we can make use of **all** of the data, while also making use of informative acoustic and prosodic features.

Adaptation for Found Data TTS

- Yamagishi et al. 2008: "Robustness of HMM-based speech synthesis." Recording condition adaptive training.
- Wan et al. 2014: "Building HMM-TTS voices on diverse data." Cluster-adaptive training to model speaker variety for speaker adaptation using mixed data.

Our experiments: Adapt to each subset of [hi, med, lo] utterances for each feature.

Frontend Features for Style Modeling

Yamagishi et al. 2003: "Modeling of various speaking styles and emotions for HMM-based speech synthesis."

- Speaking style or emotion is added as a contextual feature at the frontend and chosen at synthesis time
- Reading, Rough, Joyful, Sad

Our experiments: Add new frontend features that indicate [hi, med, lo] for each acoustic/prosodic feature.

Adding Frontend Features

```
0        1550000  x^x-sil+ax=k@  .... /E:x+       .... /J:31+19-1
1550000  2000000  x^sil-ax+k=ey@ .... /E:det+     .... /J:31+19-1
2000000  3150000  sil^ax-k+ey=p@ .... /E:content+ .... /J:31+19-1
3150000  4000000  ax^k-ey+p=k@   .... /E:content+ .... /J:31+19-1
4000000  5500000  k^ey-p+k=aa@   .... /E:content+ .... /J:31+19-1
5500000  5700000  ey^p-k+aa=d@   .... /E:content+ .... /J:31+19-1
5700000  7400000  p^k-aa+d=ax@   .... /E:content+ .... /J:31+19-1
7400000  7550000  k^aa-d+ax=t@   .... /E:content+ .... /J:31+19-1
7550000  8250000  aa^d-ax+t=er@  .... /E:content+ .... /J:31+19-1
8250000  9950000  d^ax-t+er=n@   .... /E:content+ .... /J:31+19-1
9950000  11250000 ax^t-er+n=iy@  .... /E:content+ .... /J:31+19-1
```

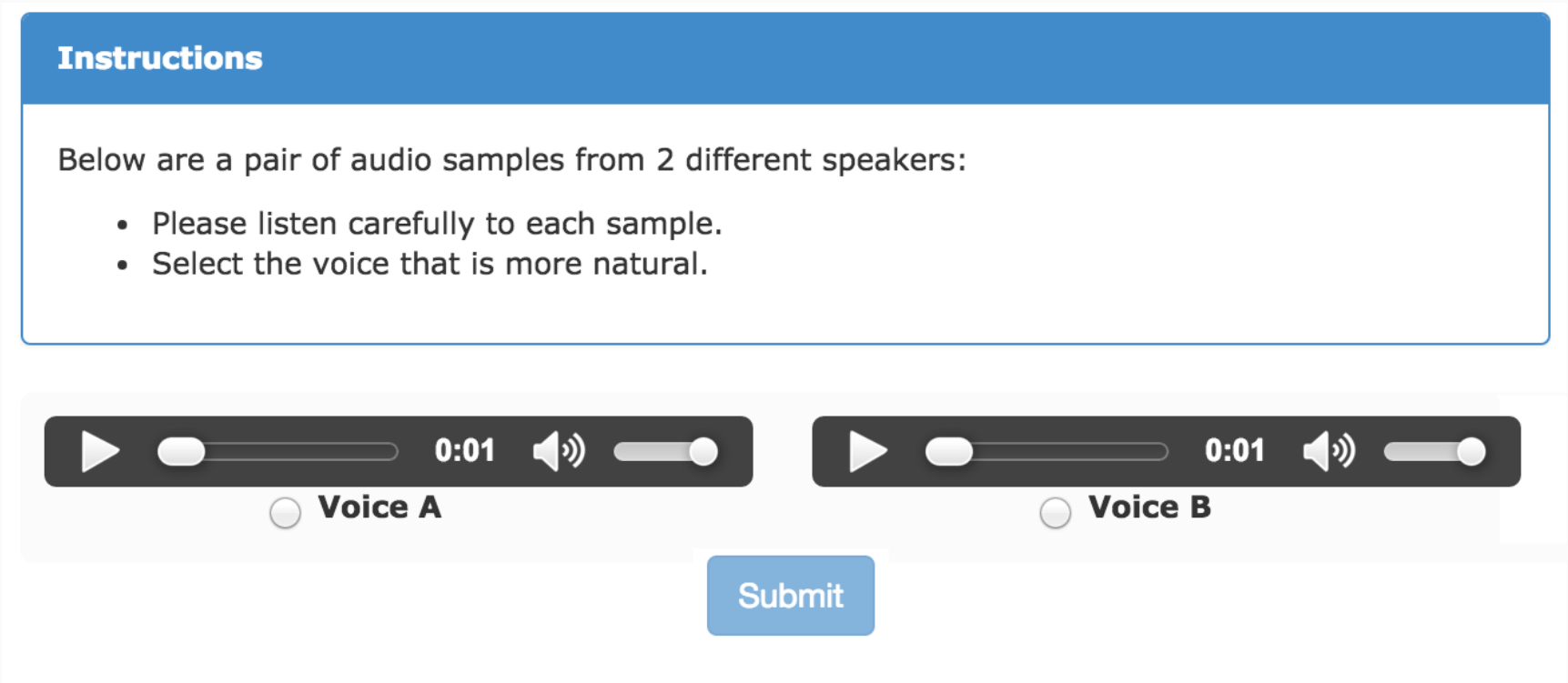
Adding Frontend Features

```
0        1550000  x^x-sil+ax=k@  .... /E:x+       .... /J:31+19-1/K:med
1550000  2000000  x^sil-ax+k=ey@ .... /E:det+     .... /J:31+19-1/K:med
2000000  3150000  sil^ax-k+ey=p@ .... /E:content+ .... /J:31+19-1/K:med
3150000  4000000  ax^k-ey+p=k@   .... /E:content+ .... /J:31+19-1/K:med
4000000  5500000  k^ey-p+k=aa@   .... /E:content+ .... /J:31+19-1/K:med
5500000  5700000  ey^p-k+aa=d@   .... /E:content+ .... /J:31+19-1/K:med
5700000  7400000  p^k-aa+d=ax@   .... /E:content+ .... /J:31+19-1/K:med
7400000  7550000  k^aa-d+ax=t@   .... /E:content+ .... /J:31+19-1/K:med
7550000  8250000  aa^d-ax+t=er@  .... /E:content+ .... /J:31+19-1/K:med
8250000  9950000  d^ax-t+er=n@   .... /E:content+ .... /J:31+19-1/K:med
9950000  11250000 ax^t-er+n=iy@  .... /E:content+ .... /J:31+19-1/K:med
```

Adding Frontend Features

```
0        1550000   x^x-sil+ax=k@   .... /E:x+        .... /J:31+19-1/K:med/L:hi
1550000  2000000   x^sil-ax+k=ey@  .... /E:det+      .... /J:31+19-1/K:med/L:hi
2000000  3150000   sil^ax-k+ey=p@  .... /E:content+  .... /J:31+19-1/K:med/L:hi
3150000  4000000   ax^k-ey+p=k@    .... /E:content+  .... /J:31+19-1/K:med/L:hi
4000000  5500000   k^ey-p+k=aa@    .... /E:content+  .... /J:31+19-1/K:med/L:hi
5500000  5700000   ey^p-k+aa=d@    .... /E:content+  .... /J:31+19-1/K:med/L:hi
5700000  7400000   p^k-aa+d=ax@    .... /E:content+  .... /J:31+19-1/K:med/L:hi
7400000  7550000   k^aa-d+ax=t@    .... /E:content+  .... /J:31+19-1/K:med/L:hi
7550000  8250000   aa^d-ax+t=er@   .... /E:content+  .... /J:31+19-1/K:med/L:hi
8250000  9950000   d^ax-t+er=n@    .... /E:content+  .... /J:31+19-1/K:med/L:hi
9950000  11250000  ax^t-er+n=iy@   .... /E:content+  .... /J:31+19-1/K:med/L:hi
```

Naturalness Evaluation: Mechanical Turk



Results

HMM Synthesis: Subset Adaptation

Percent preference for HTS voices trained adaptively using high, middle, and low partitions for each feature.

Feature	hi	med	lo
Mean f0	40.0	53.3	**56.7**
Std. dev f0	33.3	38.3	**43.3**
Mean energy	41.7	**60.0**	58.3
Std. dev energy	**43.3**	41.7	40.0
Speaking rate	**46.7**	**46.7**	35.0
Articulation	38.3	30.0	**40.0**
Duration	**40.0**	31.7	36.7

HMM Synthesis: Frontend Features

Percent preference for HTS voices trained with labels for high, medium, or low values for acoustic and prosodic features and then synthesized at each of the three settings.

Feature	hi	med	lo
Mean f0	55.0	**60.0**	51.7
Std. dev f0	60.0	55.0	**63.3**
Mean energy	48.3	**56.7**	45.0
Std. dev energy	**51.7**	50.0	**51.7**
Speaking rate	**50.0**	46.7	45.0
Articulation	**56.7**	**56.7**	**56.7**
Duration	**63.3**	50.0	56.7

HMM Synthesis: Combined Frontend Features

Percent preference for HTS voices trained with labels for multiple features.

Features	Preference
Duration (hi)	<u>63.3</u>
+ Std. dev. f0 (lo)	46.7
+ Mean f0 (med)	53.3
+ Articulation (lo)	56.7
+ Mean energy (med)	58.3
+ Std. dev. energy (lo)	<u>65.0</u>
+ Speaking rate (hi)	60.0

NN Synthesis: Subset Adaptation

Percent preference for Merlin AVM, fine-tune adapted to subsets of the data selected based on high, middle, or low values for various acoustic and prosodic features.

Feature	hi	med	lo
Mean f0	43.3	**45.0**	36.7
Std. dev f0	48.3	**60.0**	50.0
Mean energy	**53.3**	45.0	36.7
Std. dev energy	36.7	**43.3**	36.7
Speaking rate	**45.0**	**45.0**	41.7
Articulation	**50.0**	45.0	45.0
Duration	41.7	45.0	**60.0**

NN Synthesis: Frontend Features

Percent preference for Merlin voices trained on data labeled as having high, medium, or low values for features and then synthesized with each of the three settings.

Feature	hi	med	lo
Mean f0	41.7	53.3	**65.0**
Std. dev f0	51.7	**55.0**	50.0
Mean energy	46.7	48.3	**55.0**
Std. dev energy	**61.7**	50.0	60.0
Speaking rate	**50.0**	41.7	48.3
Articulation	41.7	41.7	**53.3**
Duration	48.3	**55.0**	50.0

NN Synthesis: Combined Frontend Features

Percent preference for Merlin voices trained with labels for multiple features combined.

Features	Preference
Mean f0 (lo)	<u>65.0</u>
+ Std. dev. energy (hi)	53.3
+ Duration (med)	48.3
+ Mean energy (lo)	46.7
+ Std. dev. f0 (med)	56.7
+ Articulation (lo)	35.0
+ Speaking rate (hi)	46.7

Discussion

- Adding frontend features is a more promising approach than adaptation, regardless of the type of acoustic model
- Combining frontend features generally did not lead to consistently better improvement (overfitting?)

Ongoing and Future Work

Ongoing and Future Work

- Which results generalize to other languages?
- Other types of features, such as spectral
- Try numerical values for frontend features, instead of discrete hi, med, lo
- Try different granularity than utterance-level (speaker-level, phone-level)
- Combine frontend features with adaptation
- Investigate why combining features at frontend generally does not help, and explore other ways of combining them

Acknowledgements

This work was supported by the National Science Foundation under Grant IIS 1548092.

Questions?