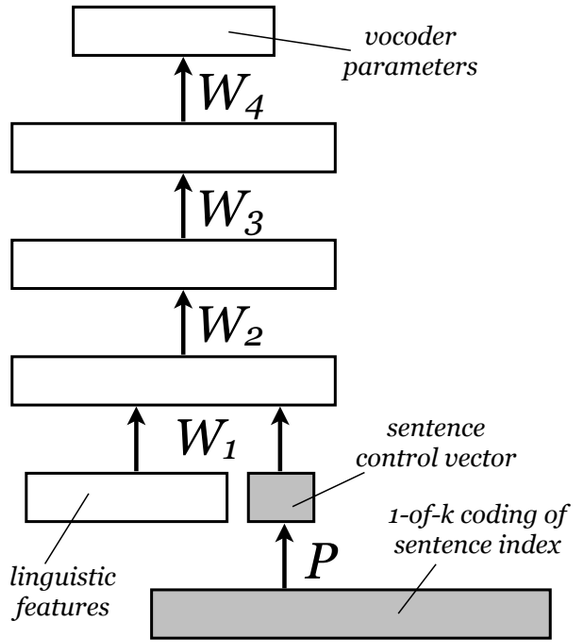
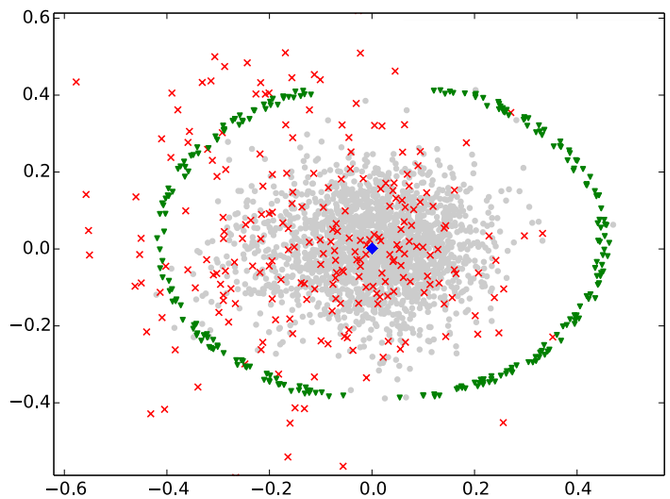
Motivation

Model



Control

Evaluation



Sentence-level control vectors for deep neural network speech synthesis

Oliver Watts
Zhizheng Wu
Simon King

Synthetic speech is boring to listen to for long periods...

Excludes sentence-level effects, which are hard to predict from text: yields synthetic speech which has acceptable sentence-level neutral rhythm and intonation

Sentences chosen for phonetic coverage and read in random order

Solution: train on data with the type of variation we would like in our synthetic speech

Difficult: no well-established way to represent sentence level effects, or predict them

Solution: learn a *'space of sentences'* automatically during training, in an unsupervised way

Established techniques for HMM synthesis: eigenvoices, MR-HSMM, CAT...

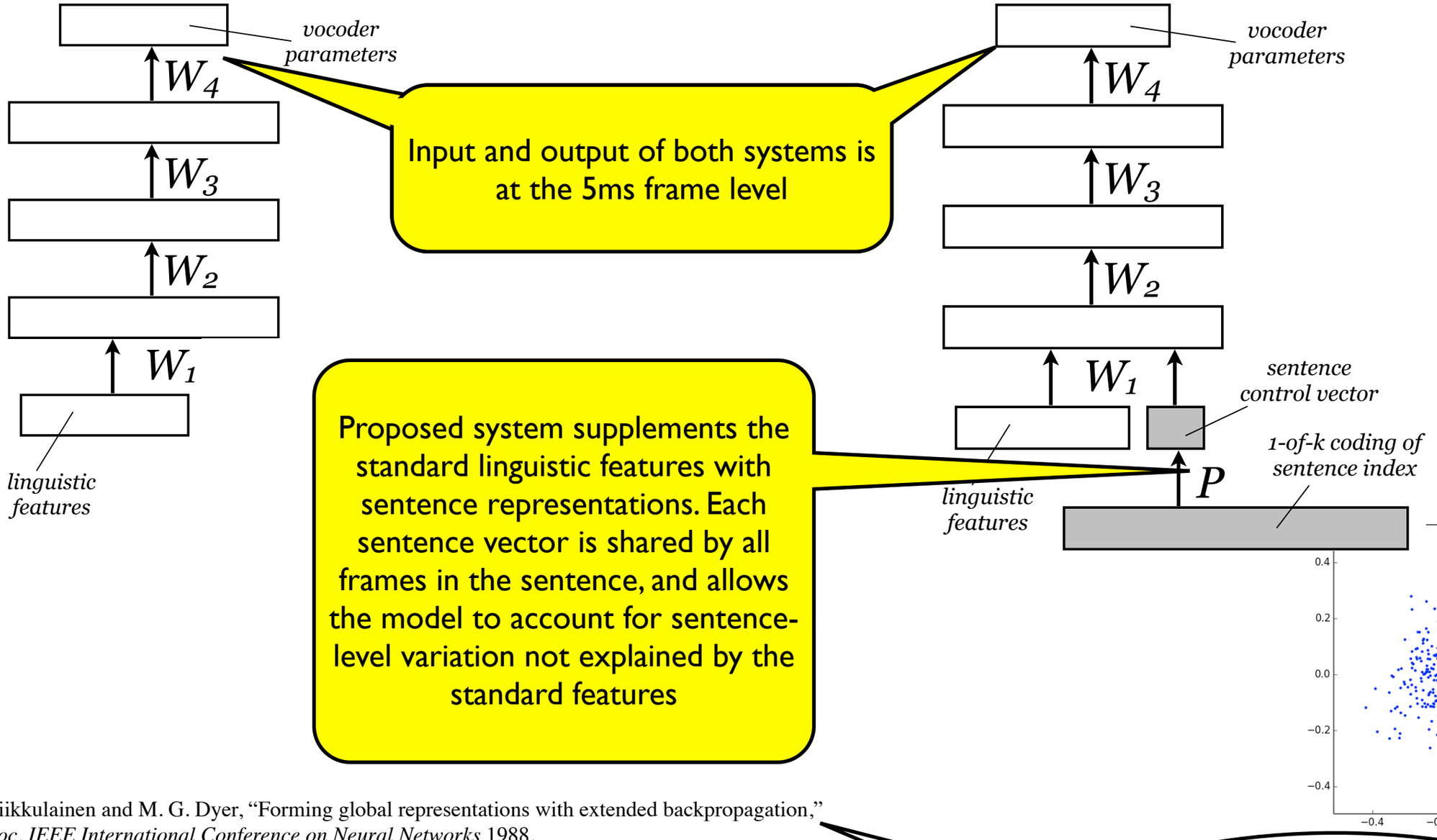
Similar technique for DNN-based synthesis?

Current focus: can we learn a useful control space? Control vectors supplied by a human operator. Future: prediction from text

The model

Conventional
DNN TTS system

Proposed
DNN TTS system



Input and output of both systems is at the 5ms frame level

Proposed system supplements the standard linguistic features with sentence representations. Each sentence vector is shared by all frames in the sentence, and allows the model to account for sentence-level variation not explained by the standard features

R. Miikkulainen and M. G. Dyer, "Forming global representations with extended backpropagation," in *Proc. IEEE International Conference on Neural Networks*,1988.

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, 2003.

K. J. Jensen and S. Riis, "Self-organizing letter code-book for text-to-phoneme neural network model," in *Interspeech*, 2000.

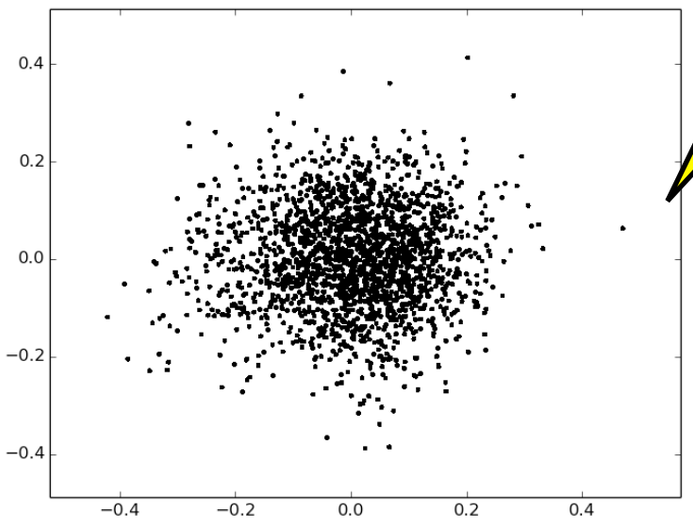
J. Bridle and S. Cox, "RecNorm: Simultaneous Normalisation and Classification applied to Speech Recognition," *NIPS*, vol. 3, 1991.

S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE TASLP*, vol. 22, no. 12, 2014.

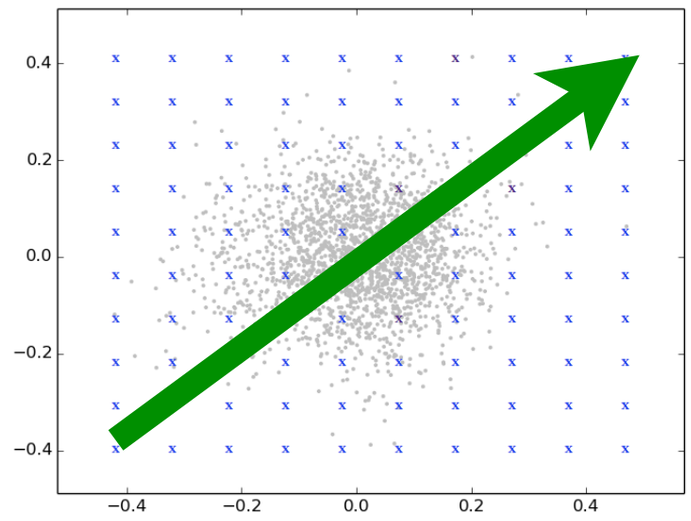
"The second goal, developing meaningful distributed representation for the data, is achieved by *extending the error signals to the input layer* and modifying the representations *as if they were weights on connections coming in to the input layer*."

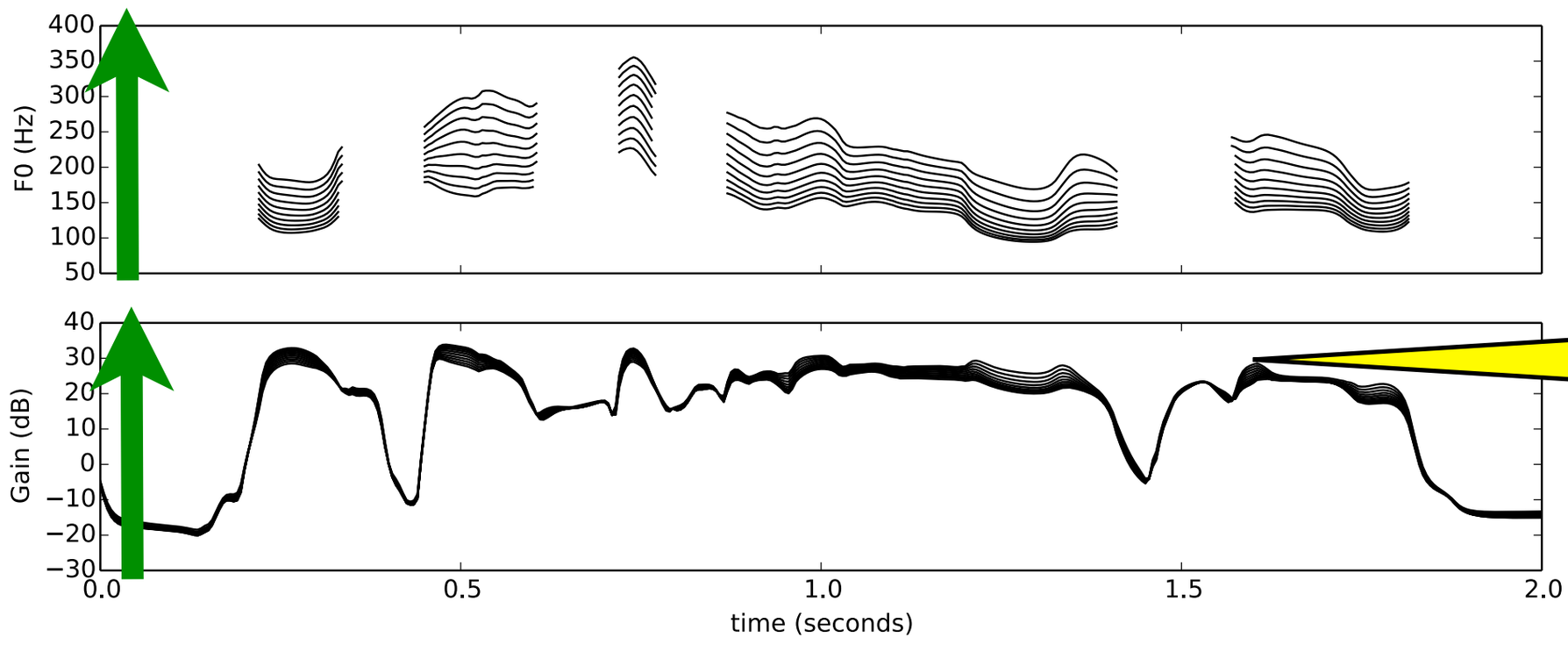
Control

The axes of the discovered space have no predefined meaning, but represent directions of sentence-level variation in the training data



The space allows control of the synthesiser's sentence-level characteristics by a human operator



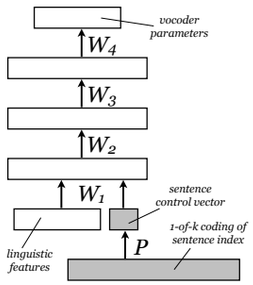


The space allows the acoustic parameters generated by the synthesiser to be modified in a way which respects their natural covariation with each other and with the linguistic context

Evaluation

Two hypotheses (evaluation without manual control):

- ‘steered’ version preferred if ideal (oracle) control vectors used
- ‘any variation better than no variation’

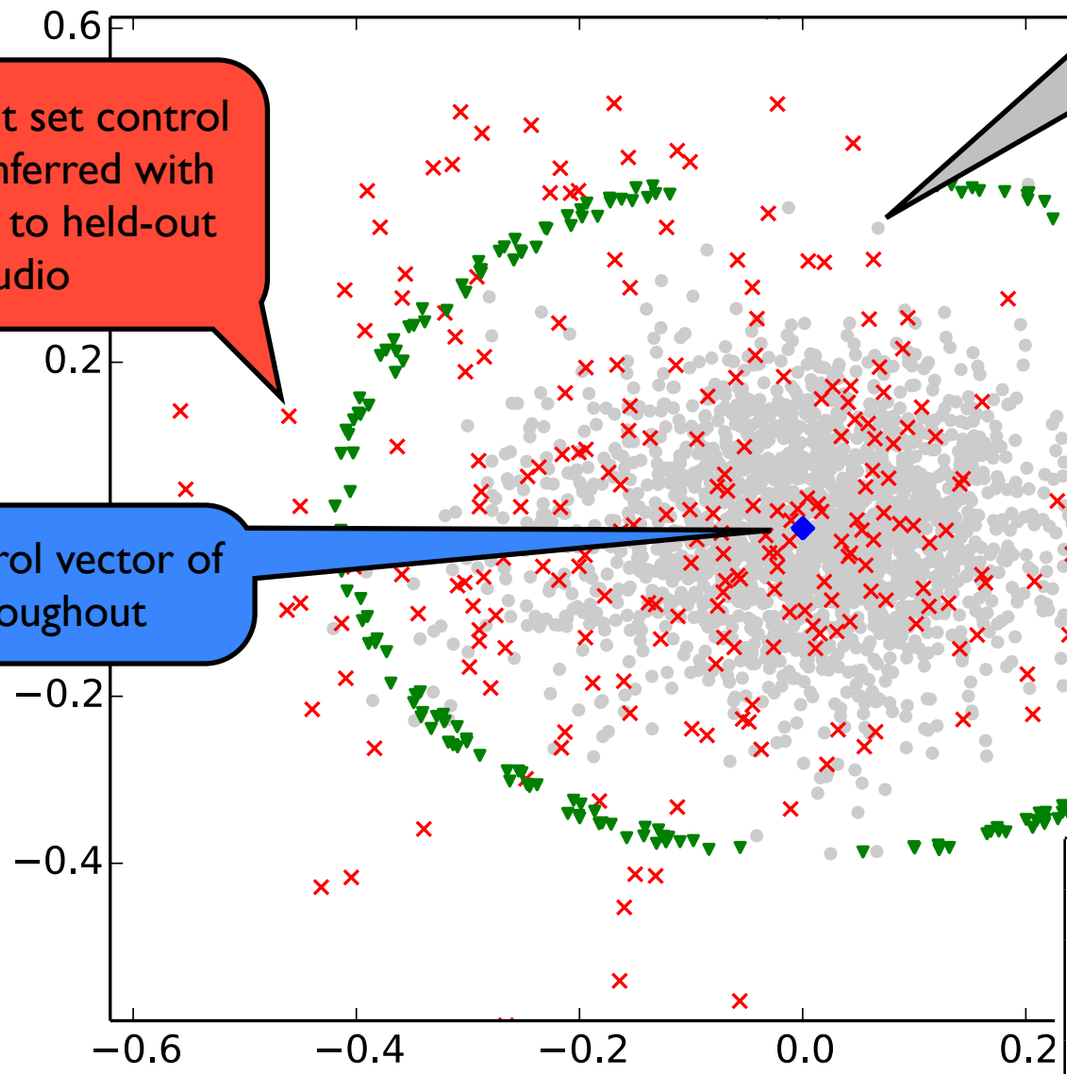


Oracle test set control vectors: inferred with reference to held-out audio

Training corpus sentences

Randomly varied control vectors: sampled uniformly from 3.8–4.0 standard deviations of the learned training corpus vectors

Fixed vector: mean control vector of training set used throughout



1	47.14	11	40.00
2	35.71	12	61.43
3	47.14	13	55.71
4	37.14	14	54.29
5	42.86	15	54.29
6	47.14	16	50.00
7	51.43	17	51.43
8	50.00	18	50.00
9	48.57	19	50.00
10	51.43	20	52.86
All	45.86	All	52.00

▼ > ◆ ✕ > ◆

3 held-out stories (70 book pages, 3.4 sentences per page) synthesised with 3 systems

Page pairs played in story order, system changed each page; within pair order of 2 competing systems is random

‘Choose the version which you would prefer to hear if you were listening to stories like this for fun’