Principles for learning controllable TTS from annotated and latent variation

Gustav Eje Henter[1], Jaime Lorenzo-Trueba[1], Xin Wang[1], Junichi Yamagishi[1,2]

[1]National Institute of Informatics, Japan *[2]The University of Edinburgh, UK*

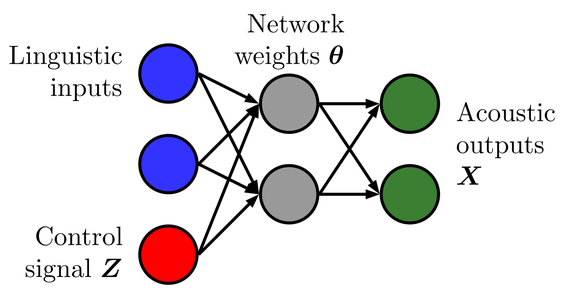
Research partially supported by the TOYOTA motor corporation

Outline

- Humans have great control over how they speak; machines do not
  - Control inputs in speech data are seldom well annotated, due to cost
- Can we learn to control also unannotated variation?
  - Yes, by joint optimisation of parameters and control inputs [1, 2, 3]
- Our main contributions:
  1. [1, 2, 3] can be interpreted as maximum likelihood with latent variables
  2. Use a sparse latent prior to learn unannotated nuances in expression
- Produces accurate emotional TTS with fine expression control

Controllable TTS

- Annotated: supervised learning
  - Baseline system
- Unannotated: heuristic joint optimisation of z and θ
  - "Discriminant condition codes" [1]
  - "Sentence-level control vectors" [2]
- Proposed method: Coarse annotation



Understanding previous heuristics

- Assume stochastic observations (speech) X depend on parameters θ and unobserved latent variables Z (unannotated control inputs) through

$$f_{X,Z}(\boldsymbol{x},\, \boldsymbol{z};\, \boldsymbol{\theta}) = f_{X\,|\,Z}\left(\boldsymbol{x}\,|\,\boldsymbol{z};\, \boldsymbol{\theta}\right) f_Z(\boldsymbol{z};\, \boldsymbol{\theta}) \tag{1}$$

- Ideal: Maximum likelihood and maximum a-posteriori (MAP) estimation

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \ln \int f_{X,Z}(\boldsymbol{x},\, \boldsymbol{z};\, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{z} \tag{2}$$

$$\widehat{\boldsymbol{z}}_{\mathrm{MAP}}(\boldsymbol{x};\, \boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{z}} \ln \left(f_{X\,|\,Z}\left(\boldsymbol{x}\,|\,\boldsymbol{z};\, \boldsymbol{\theta}\right) f_Z(\boldsymbol{z};\, \boldsymbol{\theta})\right) \tag{3}$$

- Heuristic joint point-estimation objective used in [1, 2, 3]

$$\left(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}),\, \widehat{\boldsymbol{z}}(\boldsymbol{x})\right) = \operatorname*{argmax}_{(\boldsymbol{\theta},\, \boldsymbol{z})} \ln f_{X\,|\,Z}\left(\boldsymbol{x}\,|\,\boldsymbol{z};\, \boldsymbol{\theta}\right) \tag{4}$$

  - $f_{X\,|\,Z}\left(\boldsymbol{x}\,|\,\boldsymbol{z};\, \boldsymbol{\theta}\right)$ is a DNN/RNN with input $\boldsymbol{z}$, weights $\boldsymbol{\theta}$, and output $\mathbb{E}(\boldsymbol{x})$
  - $f_{X\,|\,Z}$ isotropic Gaussian $\Rightarrow$ minimum mean-squared error (MMSE) estimation
- We prove that

$$\left(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}),\, \widehat{\boldsymbol{z}}(\boldsymbol{x})\right) \approx \left(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{x}),\, \widehat{\boldsymbol{z}}_{\mathrm{MAP}}(\boldsymbol{x};\, \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{x}))\right) \tag{5}$$

  (in a local, not global sense), assuming
  - Sharp posterior: $f_{X\,|\,Z}\left(\boldsymbol{x}\,|\,\boldsymbol{z};\, \boldsymbol{\theta}\right) \approx 0$ unless $\boldsymbol{z} \approx \widehat{\boldsymbol{z}}_{\mathrm{MAP}}(\boldsymbol{x};\, \boldsymbol{\theta})$
  - Flat prior: $f_Z(\boldsymbol{z};\, \boldsymbol{\theta}) \approx c$ on the support of the distribution
- The proof is based on the EM-algorithm; Jensen's inequality is a key step
- Interpretation:
  - The established heuristic approximates maximum-likelihood estimation
  - Estimated input values $\widehat{\boldsymbol{z}}(\boldsymbol{x})$ are "poor man's latent variables"
- Benefits:
  - We now understand the goals and approximations of established methods
  - We can use latent-variable theory to improve controllable TTS

Experiment setup

Can we learn nuances in emotional expression from coarsely annotated speech?

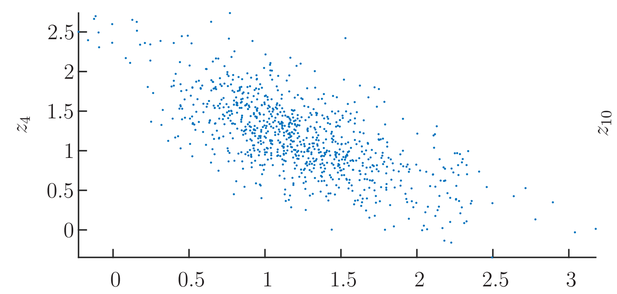
- Idea: Learn a single emotional space with a sparse prior to separate emotions
- Data: Japanese-language single-speaker emotional speech database
  - 7 emotions (acted), 1200 utterances each (80% used for training)
  - Tried to keep emotional expression as consistent as possible
  - 8400 total utterances, more than 1000 minutes
- Features: Open JTalk linguistic features; WORLD vocoder with MLPG
  - Alignment and duration prediction using emotion-aware HSMM (HTS)
- Deep RNN: Code from CURRENNT toolkit; MMSE training using SGD

1. Baseline system (Ba): Fixed, one-hot emotion input
2. Proposed system (P): 14-dimensional learned latent space
   - Emotional expression $\widehat{\boldsymbol{z}}^{(n)} \in \mathbb{R}^{14}$ assumed constant for each utterance n, to encourage learned control parameters to represent utterance-level variation
   - Each emotion occupies a separate, orthogonal 2D subspace
   - Example: sparsity pattern $\widehat{\boldsymbol{z}}^{(n)} = \boxed{0\,0 \mid \widehat{\boldsymbol{z}}_3^{(n)}\, \widehat{\boldsymbol{z}}_4^{(n)} \mid 0\,0\,\ldots\,0\,0}$ when n belongs to emotion two; use "two-hot initialisation"
   - Trained using constrained heuristic objective ("poor man's latent variables")

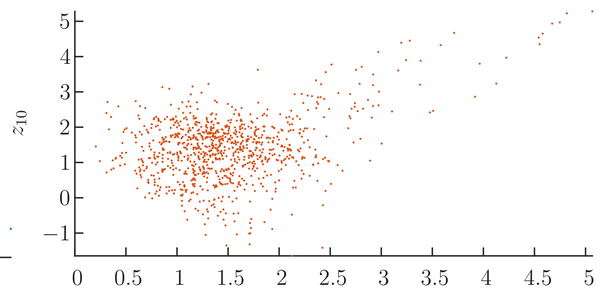
Examples of learned control vectors

- Scatterplots of learned control vectors $\{\widehat{\boldsymbol{z}}\}$ in different emotional subspaces

Happy:



Sad:



Emotion recognisability

- Verify that proposal maintains the recognisability of synthesised emotions
- Generate speech stimuli from:
  - One-hot baseline emotional TTS (discrete $\widehat{\boldsymbol{z}}^{(n)}$, not learned)
  - Proposed system (learned, cont. $\widehat{\boldsymbol{z}}^{(n)}$) using per-emotion mean control input
  - Also test held-out natural speech recordings (N)
- Emotion classification test
  - Listeners classify stimuli into the seven emotional categories or "other"
  - 75 crowdsourced listeners, 1162 responses
  - Italicised Bonferroni-corrected p-values are not significant at level $\alpha = 0.05$

	% correct			p-values	
Emotion	Natural (N)	Baseline (Ba)	Proposed (P)	N vs. P	Ba vs. P
Neutral	88	69	86	*1.000*	*0.345*
Happy	95	85	88	*1.000*	*1.000*
Calm	71	63	46	*0.057*	*0.576*
Excited	32	28	18	*0.576*	*1.000*
Sad	93	72	70	0.045	*1.000*
Insecure	71	61	59	*0.863*	*1.000*
Angry	91	91	93	*1.000*	*1.000*
All	77	67	66	0.004	*1.000*

Emotional nuance discrimination

- Verify that changing the control vector alters the emotional expression
- Use P to generate pairs of same-sentence stimuli with different z-vectors
- Rand: Pairs of distinct latent vectors randomly drawn from $\{\widehat{\boldsymbol{z}}\}$
- Far: Latent vectors taken from the pairs furthest in Mahalanobis distance
- ABX listening test
  - A, B, X stimuli based on the above stimulus pairs
  - "Which of A and B has the most similar emotional expression to X?"
  - 18 crowdsourced listeners, 950 responses

	% correct		p-values	
Emotion	Rand	Far	Rand	Far
Neutral	74	70	$<10^{-3}$	0.003
Happy	66	91	0.027	$<10^{-12}$
Calm	61	90	*0.149*	$<10^{-11}$
Excited	57	83	*0.364*	$<10^{-7}$
Sad	80	100	$<10^{-5}$	≈ 0
Insecure	57	88	*0.364*	$<10^{-10}$
Angry	80	99	$<10^{-5}$	$<10^{-19}$
All	68	89	$<10^{-13}$	$<10^{-71}$

Conclusions

- Proposal maintains emotion recognition rate of baseline synthesiser
  - TTS emotion recognition only moderately worse than for natural speech
- Latent space control creates audible differences in emotional expression
  - Noticeable differences are mostly differences in emotional strength, even though the recordings tried to keep emotional strength constant
  - Sad speech shows lower TTS quality for z-values in the point-cloud tail
- Further exploiting connections to latent-variable theory is future work

References

[1] S. Xue, O. Abdel-Hamid, H. Jiang, L.-R. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," IEEE/ACM T. Audio Speech, vol. 22, no. 12, pp. 1713–1725, 2014.

[2] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in Proc. Interspeech, 2015, pp. 2217–2221.

[3] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in Proc. ICASSP, 2017, pp. 4905–4909.