Multilingual Text-to-Speech

601.764

4/20/23

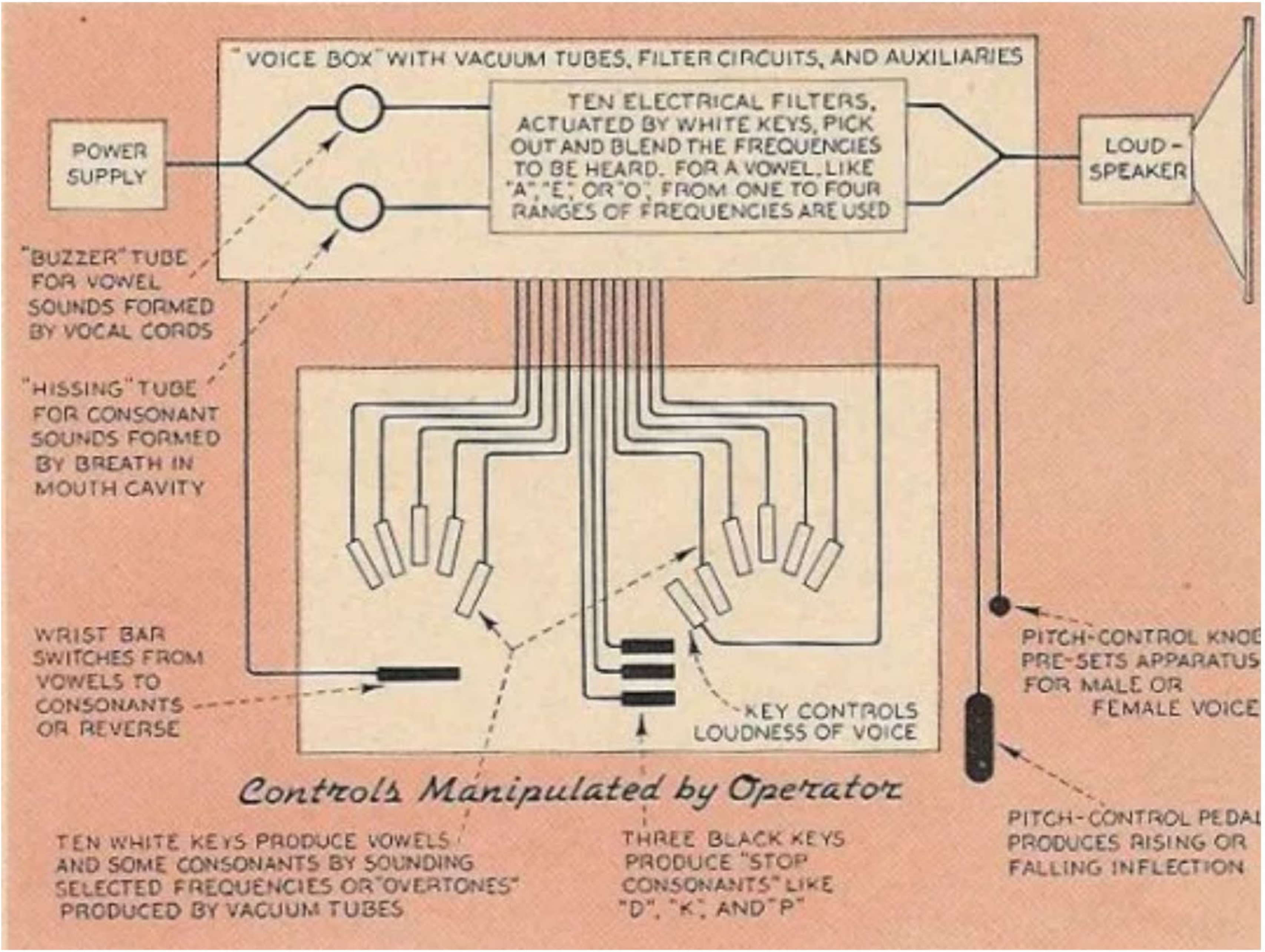
Homer Dudley’s Voder 1940

https://120years.net/the-voder-vocoderhomer-dudleyusa1940/

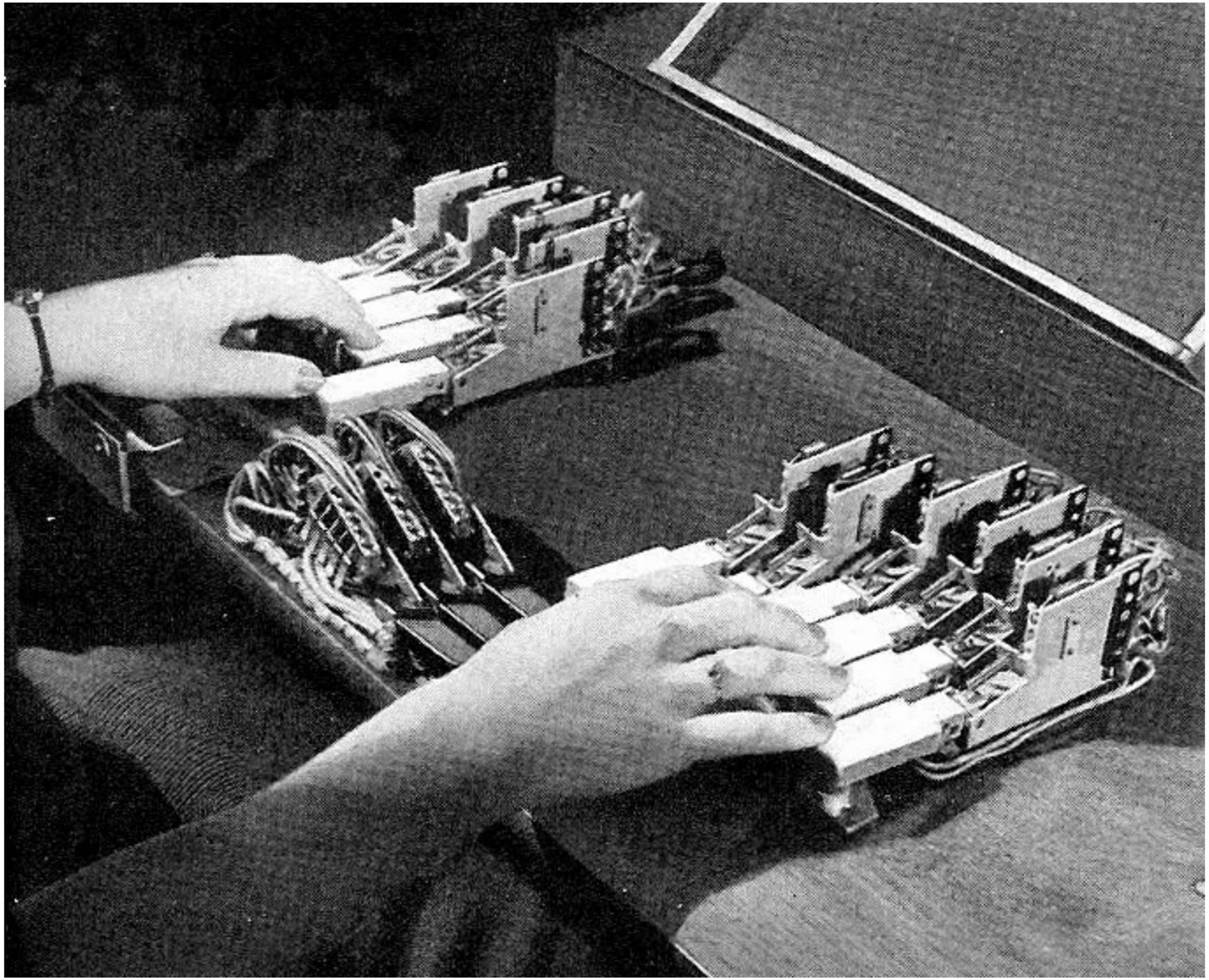
“The Voder was first unveiled in 1939 at the New York World Fair (where it was demonstrated at hourly intervals) and later in 1940 in San Francisco. There were twenty trained operators known as the ‘girls’ who handled the machine much like a musical instrument such as a piano or an organ, but they managed to successfully produce human speech during the demonstrations. In the New York Fair demonstration, which was repeated frequently, the announcer gave a simple running discussion of the circuit to which the girl operator replied through the Voder. This was done by manipulating fourteen keys with the fingers, a bar with the left wrist and a foot pedal with the right foot.”

https://120years.net/the-voder-vocoderhomer-dudleyusa1940/

Voder at the world fair



Voder diagram

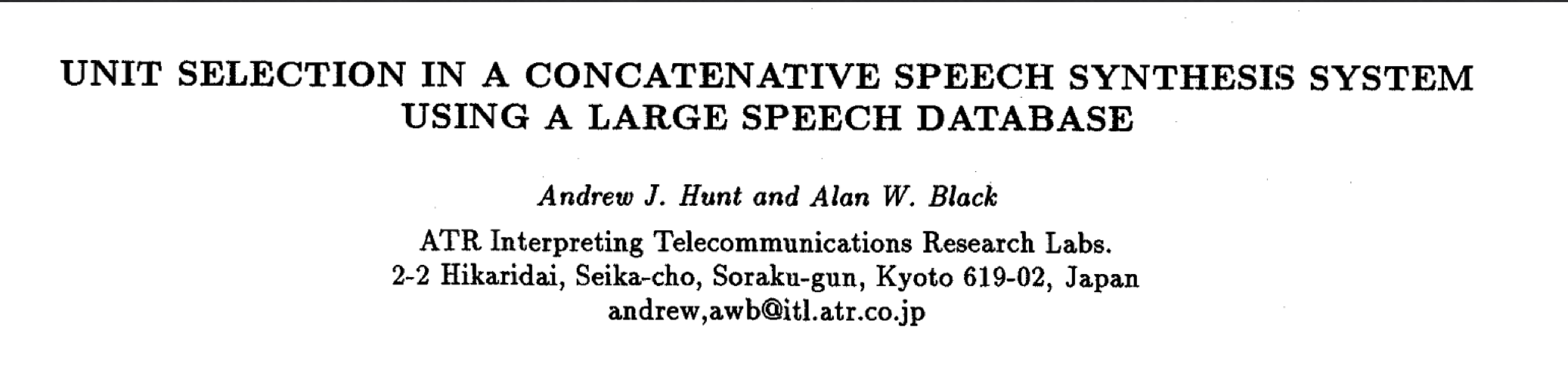

Voder keyboard and wrist controls

“The Voder was outwardly similar to a parlor organ. The white keys produced vowels; the black keys acted as “stop” consonants (such as *t* and *d*), cutting off airflow; and a foot pedal changed the pitch.”

https://muse.jhu.edu/article/491050

Still to this day…

1996



UNIT SELECTION IN A CONCATENATIVE SPEECH SYNTHESIS SYSTEM USING A LARGE SPEECH DATABASE

Andrew J. Hunt and Alan W. Black

ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
andrew,awb@itl.atr.co.jp

- Select and concatenate units from a large database
- Transition network similar to HMMs
- … Experiments

“Both training methods have been applied to a range of synthesis databases including Japanese and English, and male and female speech. Synthesized speech produced from weights of either training method is consistently better than that produced with hand-tuned weights. However, hand tuning of global unit selection parameters can improve the quality of synthesis with automatically trained weights”

2009

Review

Statistical parametric speech synthesis

Heiga Zen[a,b,*], Keiichi Tokuda[a], Alan W. Black[c]

[a] *Department of Computer Science and Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan*
[b] *Cambridge Research Laboratory, Toshiba Research Europe Ltd., 208 Cambridge Science Park, Milton Road, Cambridge CB4 0GZ, UK*
[c] *Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*

Received 14 January 2009; received in revised form 6 April 2009; accepted 8 April 2009

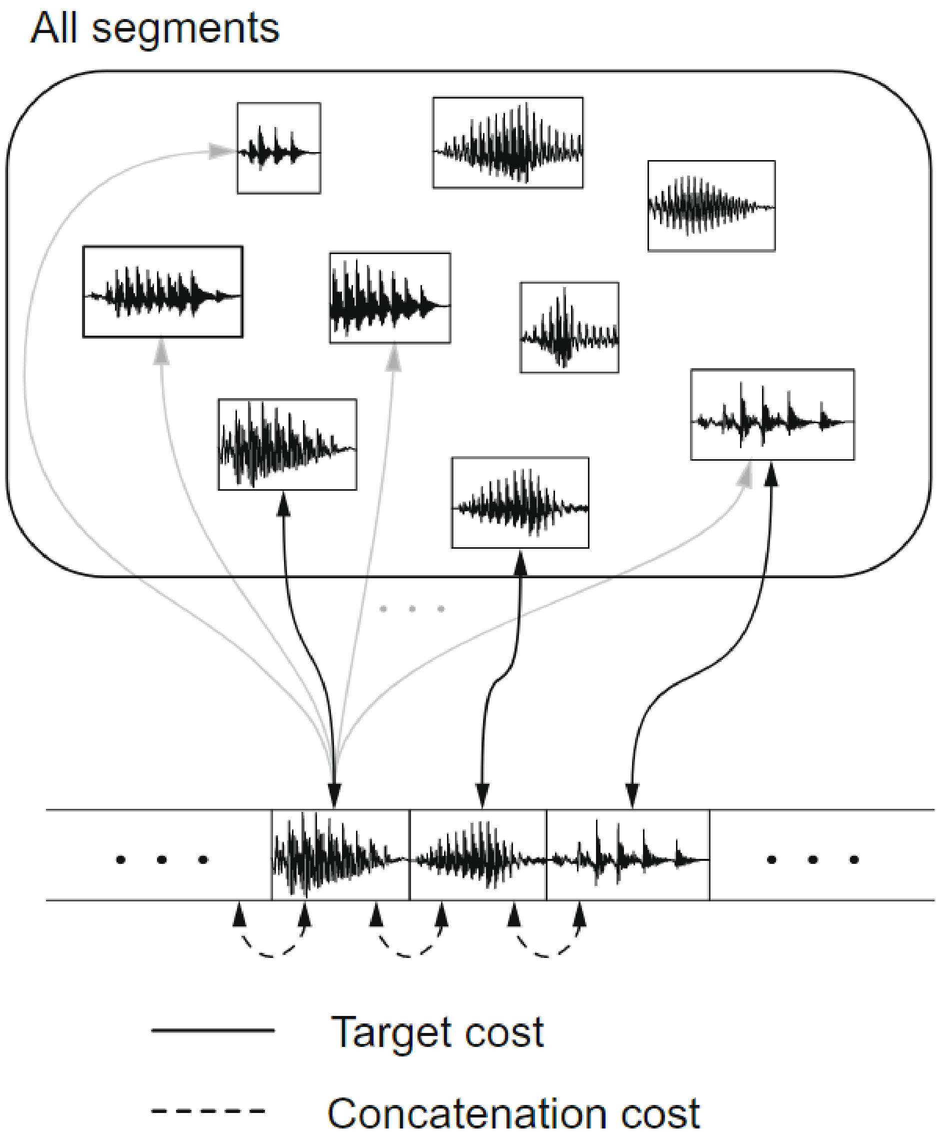


Fig. 1. Overview of general unit-selection scheme. Solid lines represent target costs and dashed lines represent concatenation costs.

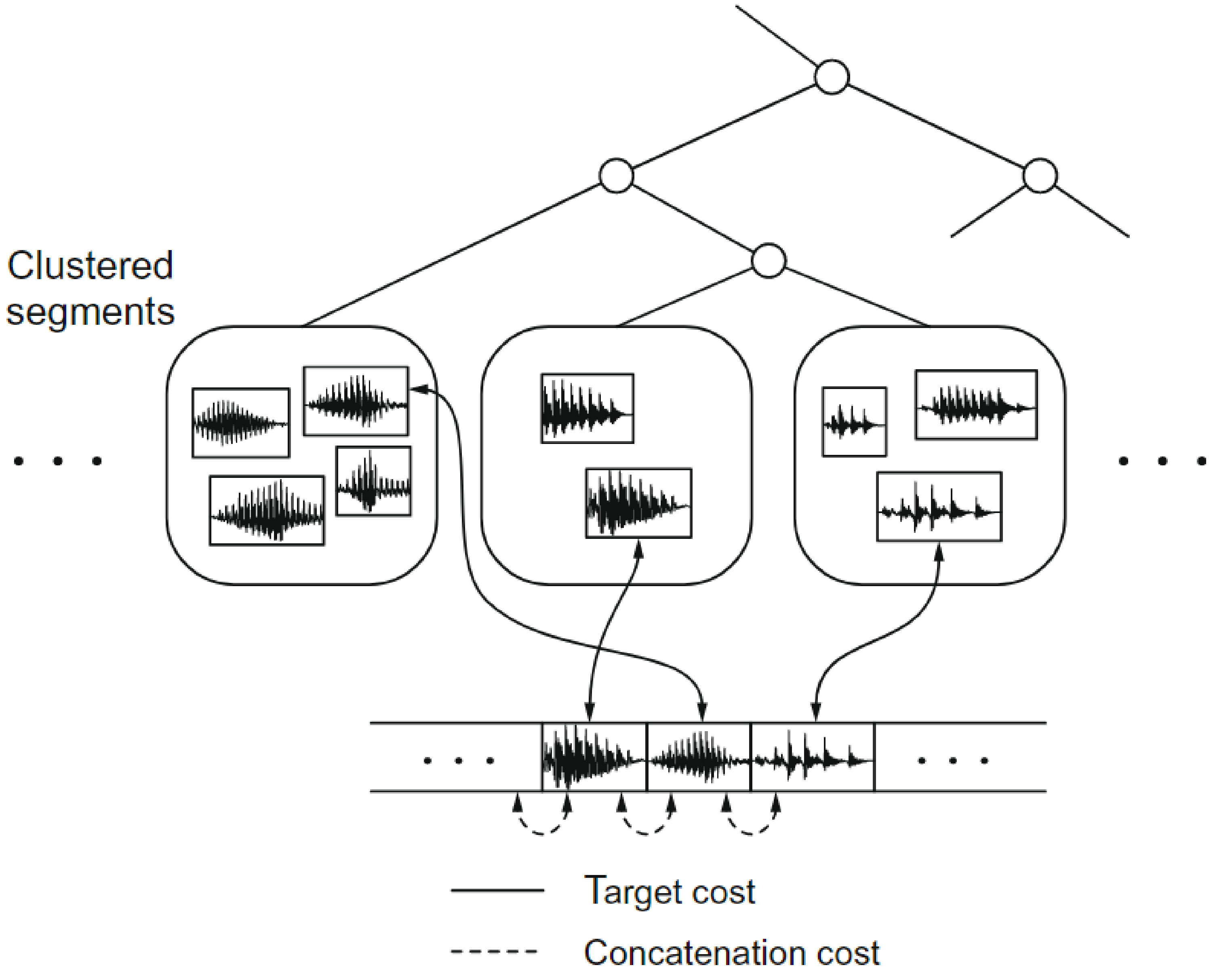


Fig. 2. Overview of clustering-based unit-selection scheme. Solid lines represent target costs and dashed lines represent concatenation costs.

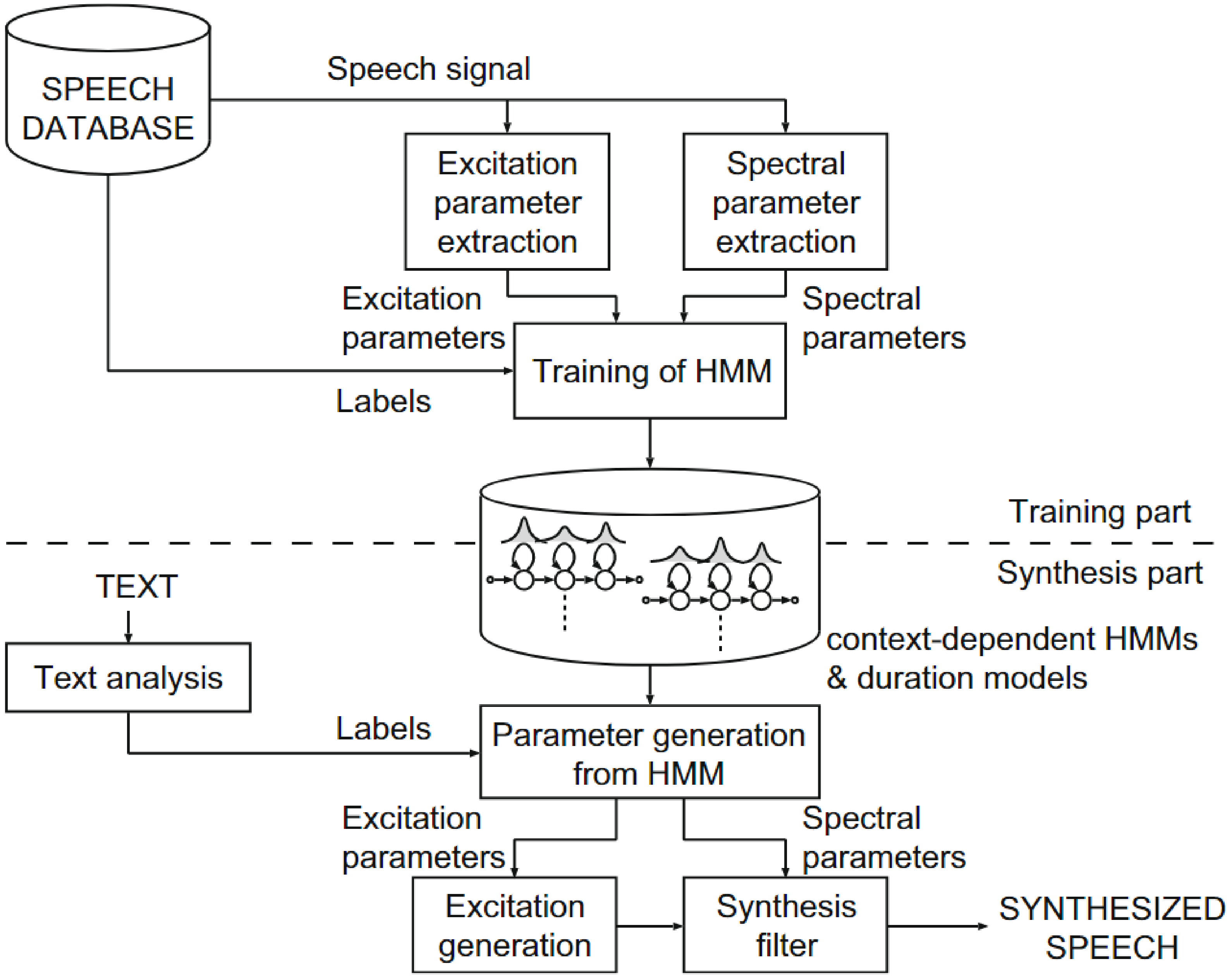


Fig. 3. Block-diagram of HMM-based speech synthesis system (HTS).

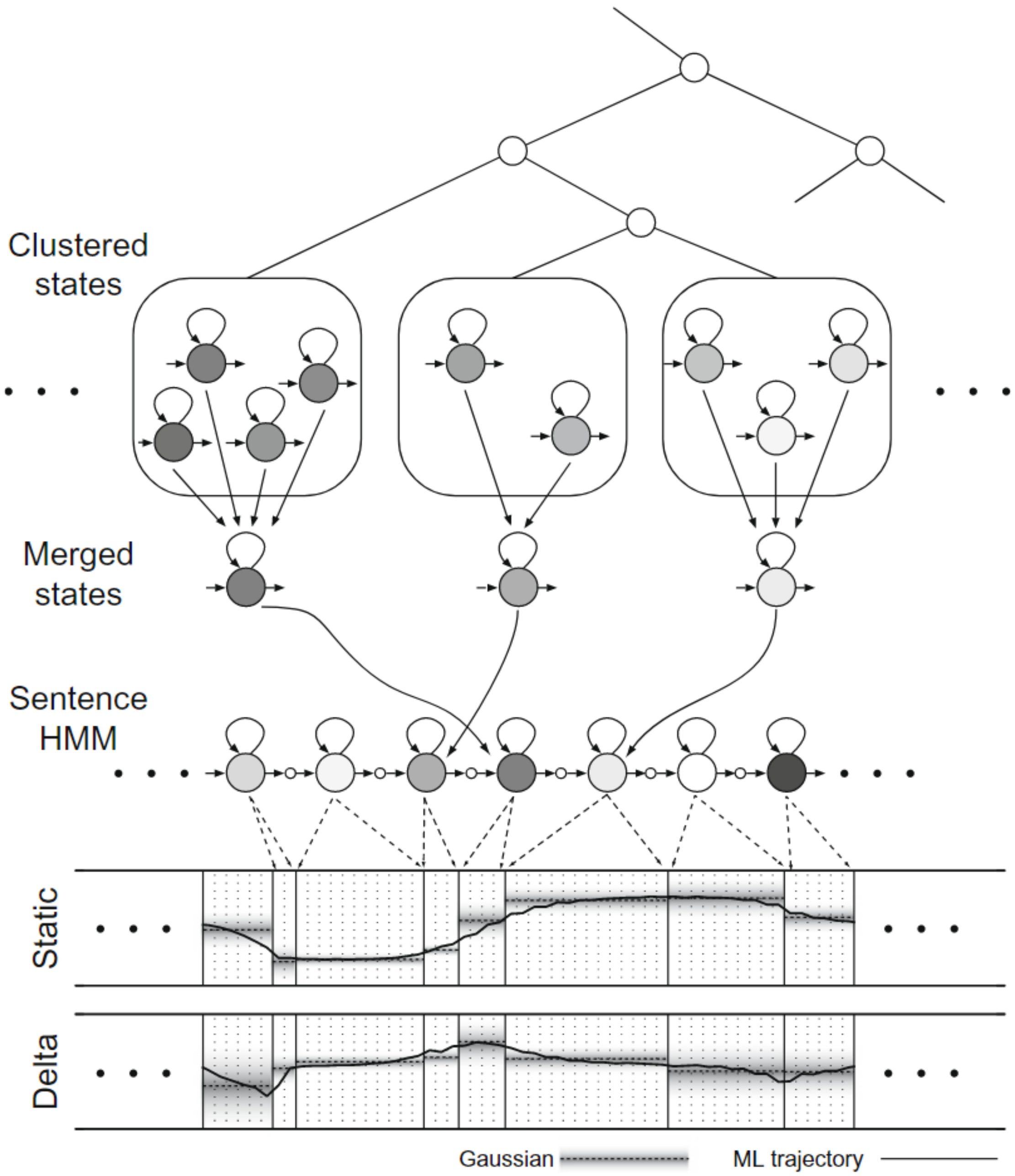


Fig. 5. Overview of HMM-based speech synthesis scheme.

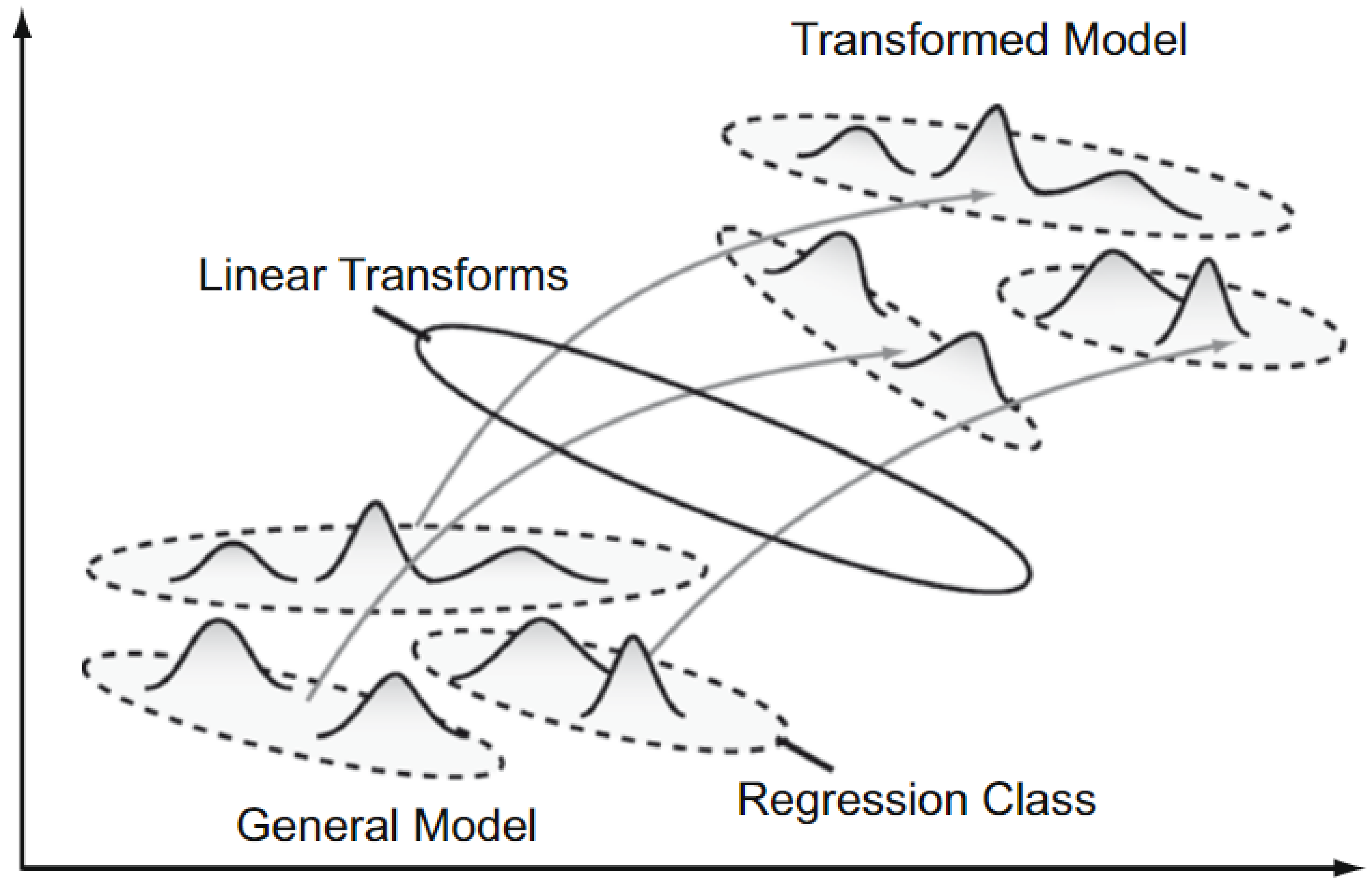


Fig. 6. Overview of linear-transformation-based adaptation technique.

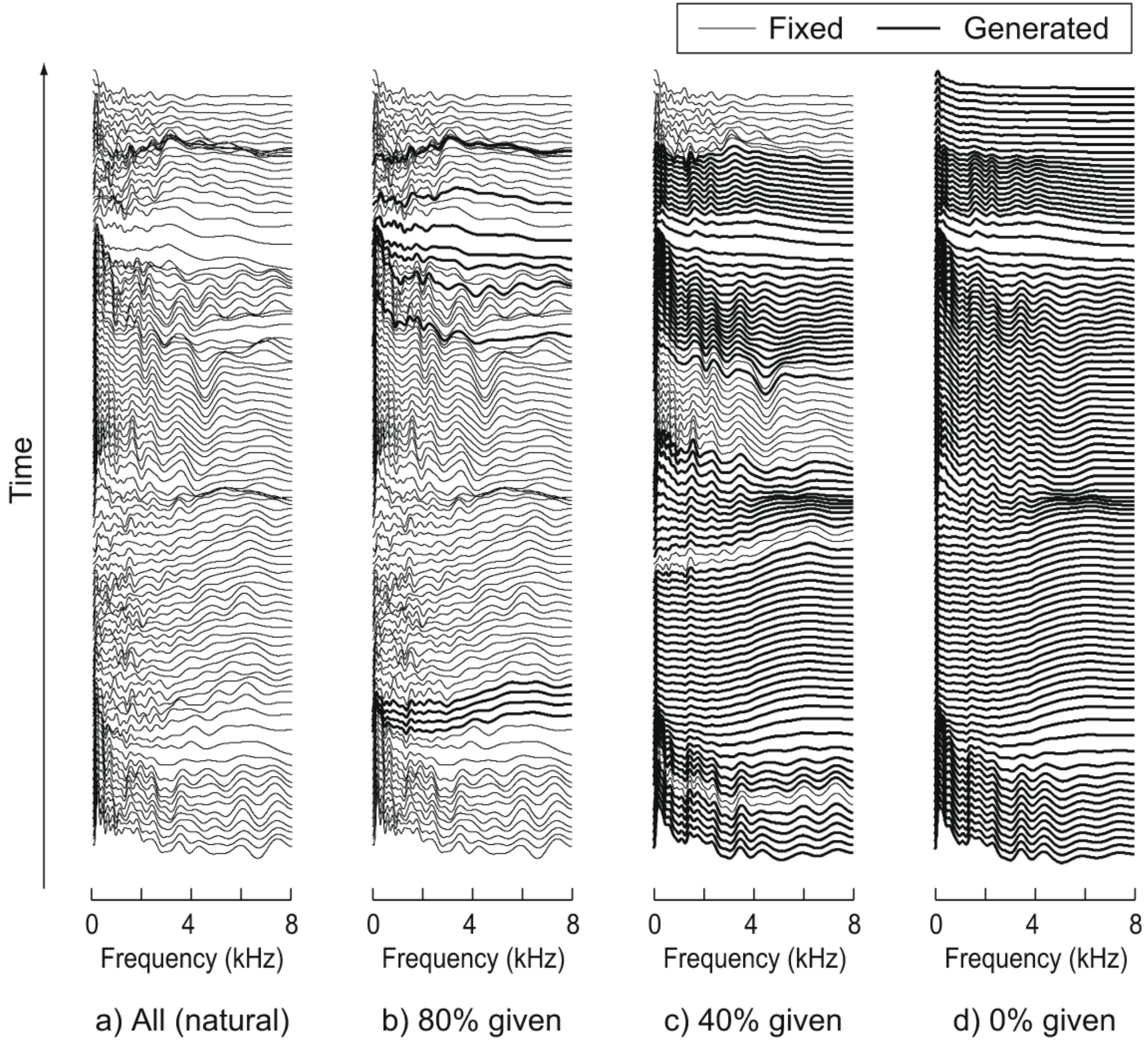


Fig. 16. Spectra generated by conditional parameter generation algorithm. Here (a) all, (b) 80%, (c) 40%, and (d) no frames are given to conditional parameter generation algorithm. Thin lines indicate given frames and thick lines indicate those generated.

2016

WaveNet: A Generative Model for Raw Audio

Aäron van den Oord **Sander Dieleman** **Heiga Zen**[†]

Karen Simonyan **Oriol Vinyals** **Alex Graves**

Nal Kalchbrenner **Andrew Senior** **Koray Kavukcuoglu**

{avdnoord, sedielem, heigazen, simonyan, vinyals, gravesa, nalk, andrewsenior, korayk}@google.com
Google DeepMind, London, UK
[†] Google, London, UK

WaveNet Examples

- https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio

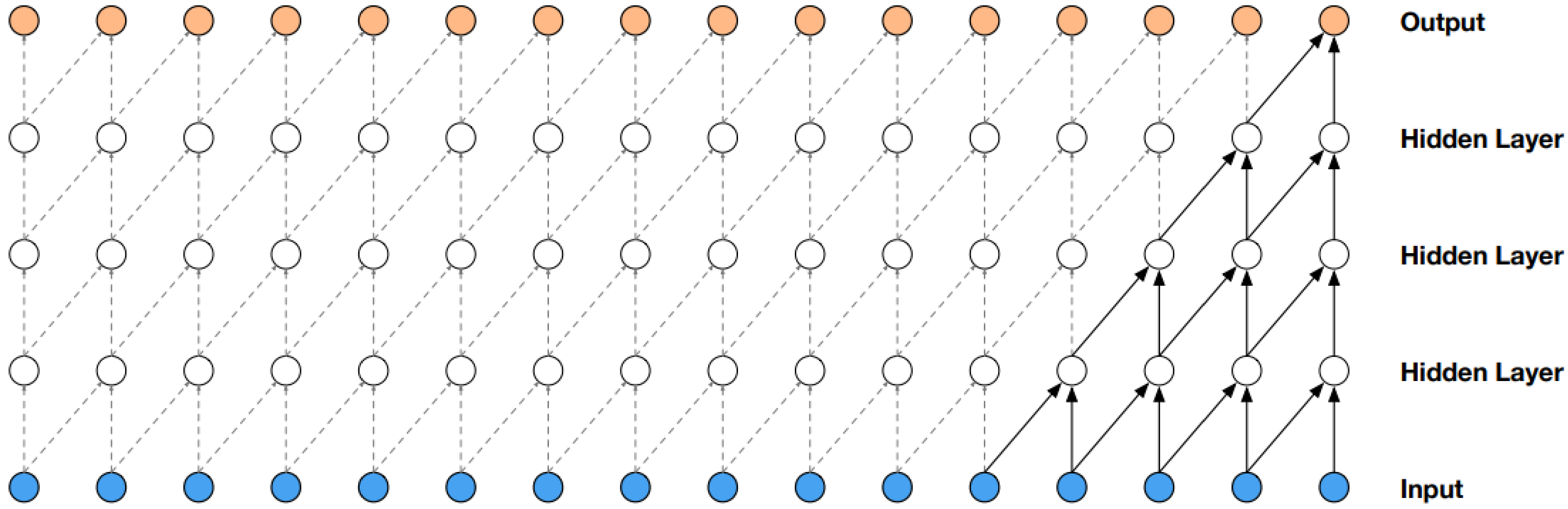


Figure 2: Visualization of a stack of causal convolutional layers.

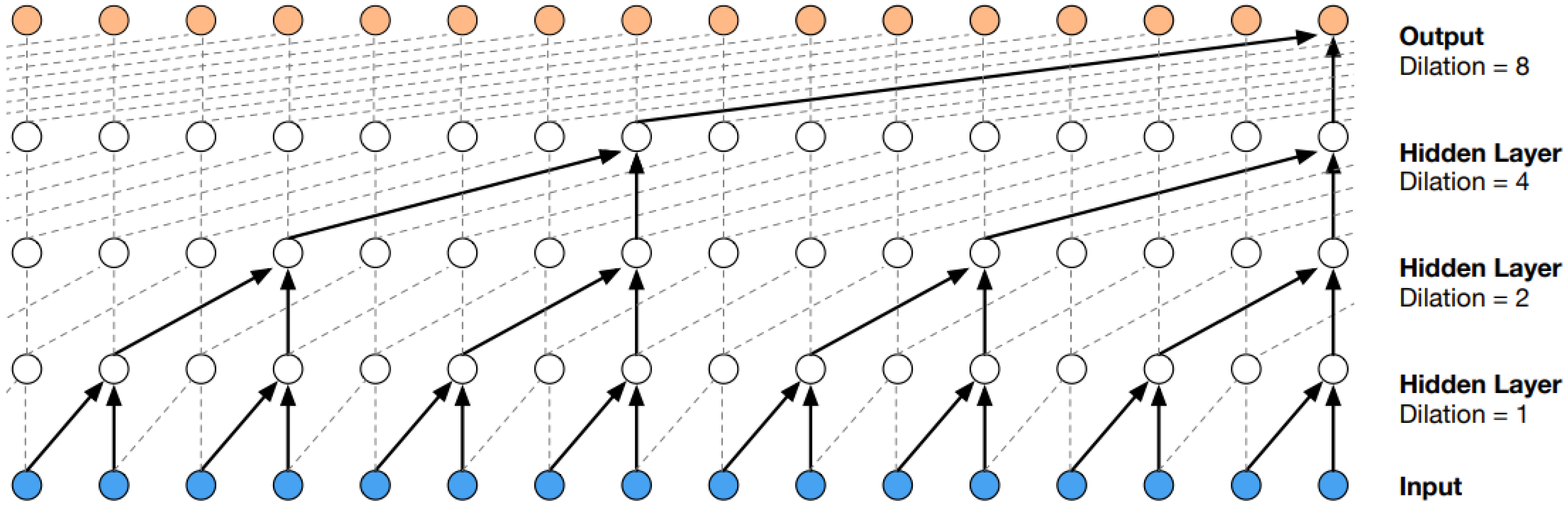
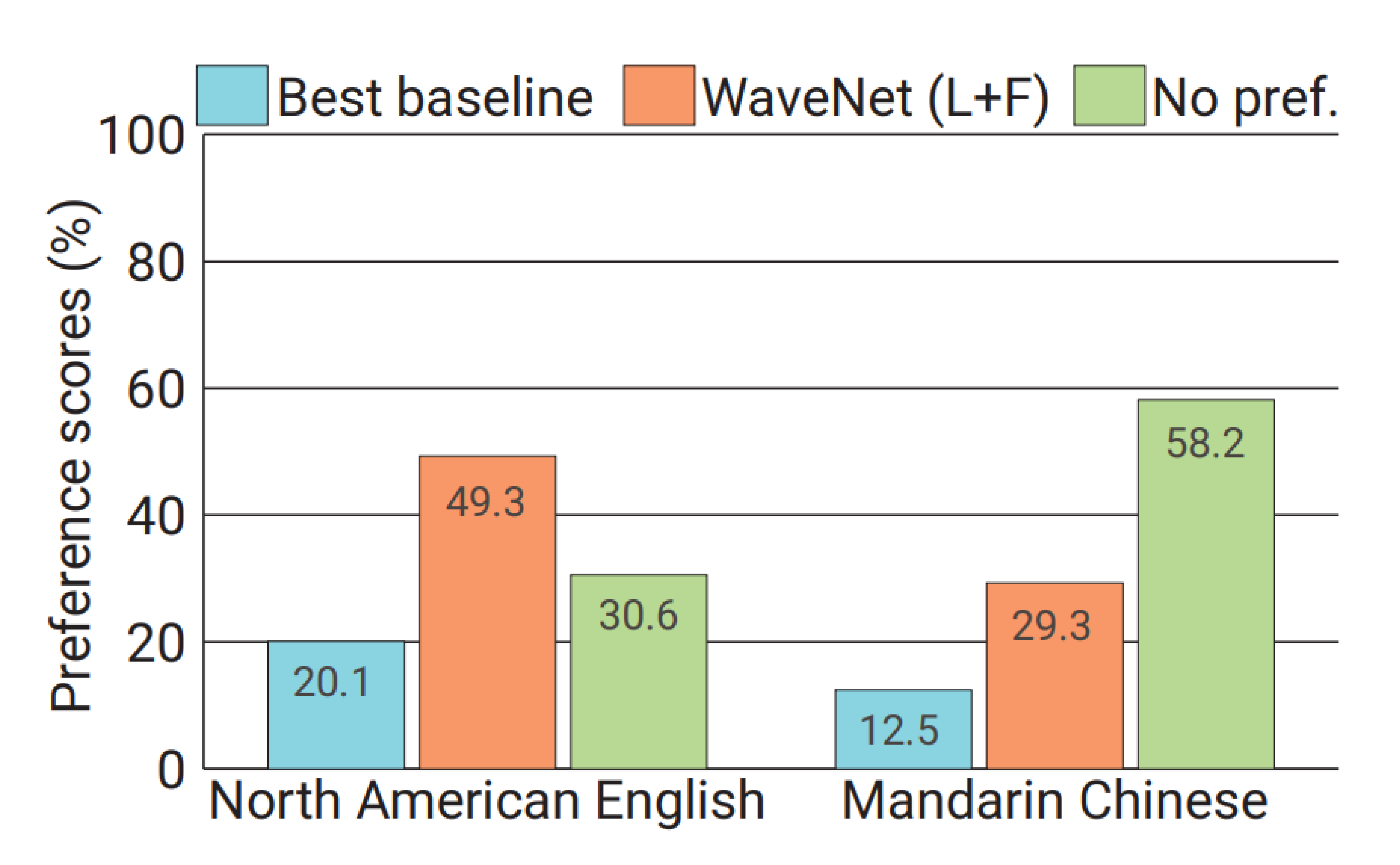


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

Table 1: Subjective 5-scale mean opinion scores of speech samples from LSTM-RNN-based statistical parametric, HMM-driven unit selection concatenative, and proposed WaveNet-based speech synthesizers, 8-bit μ-law encoded natural speech, and 16-bit linear pulse-code modulation (PCM) natural speech. WaveNet improved the previous state of the art significantly, reducing the gap between natural speech and best previous model by more than 50%.



Best baseline
WaveNet (L+F)
No pref.
Preference scores (%)
100
80
60
40
20
0
20.1
49.3
30.6
12.5
29.3
58.2
North American English
Mandarin Chinese

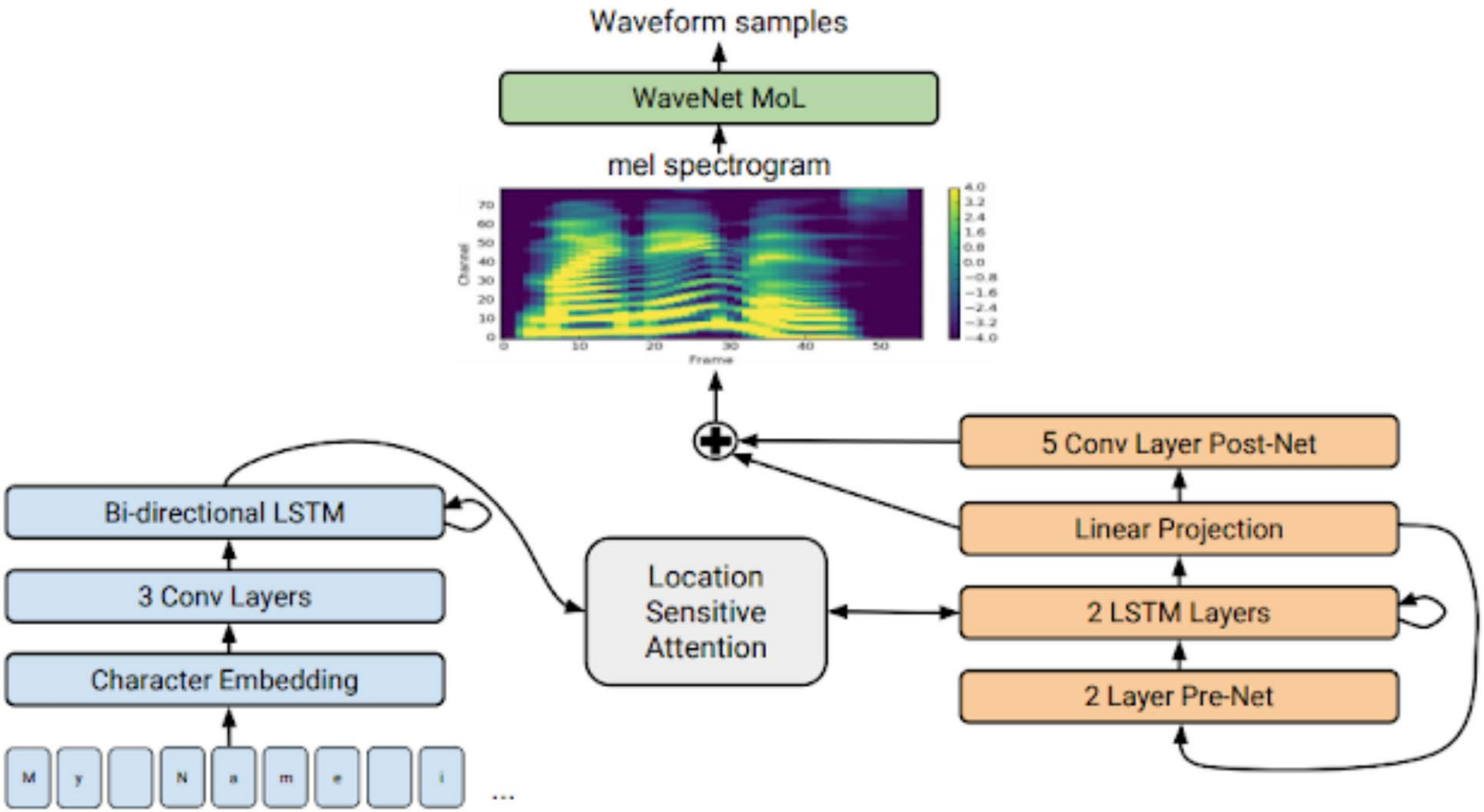
2017

NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

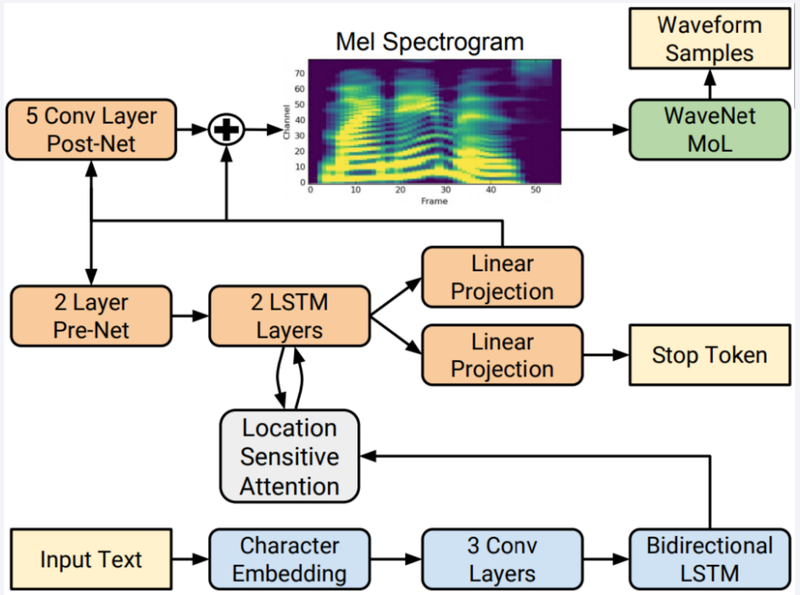
Jonathan Shen[1], *Ruoming Pang*[1], *Ron J. Weiss*[1], *Mike Schuster*[1], *Navdeep Jaitly*[1], *Zongheng Yang*[*2], *Zhifeng Chen*[1], *Yu Zhang*[1], *Yuxuan Wang*[1], *RJ Skerry-Ryan*[1], *Rif A. Saurous*[1], *Yannis Agiomyrgiannakis*[1], *and Yonghui Wu*[1]

[1]Google, Inc., [2]University of California, Berkeley,
{jonathanasdf,rpang,yonghui}@google.com

Tactotron2



A detailed look at Tacotron 2's model architecture. The lower half of the image describes the sequence-to-sequence model that maps a sequence of letters to a spectrogram. For technical details, please refer to [the paper](#).



Model Description

The Tacotron 2 and WaveGlow model form a text-to-speech system that enables user to synthesise a natural sounding speech from raw transcripts without any additional prosody information. The Tacotron 2 model produces mel spectrograms from input text using encoder-decoder architecture. WaveGlow (also available via torch.hub) is a flow-based model that consumes the mel spectrograms to generate speech.

This implementation of Tacotron 2 model differs from the model described in the paper. Our implementation uses Dropout instead of Zoneout to regularize the LSTM layers.

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

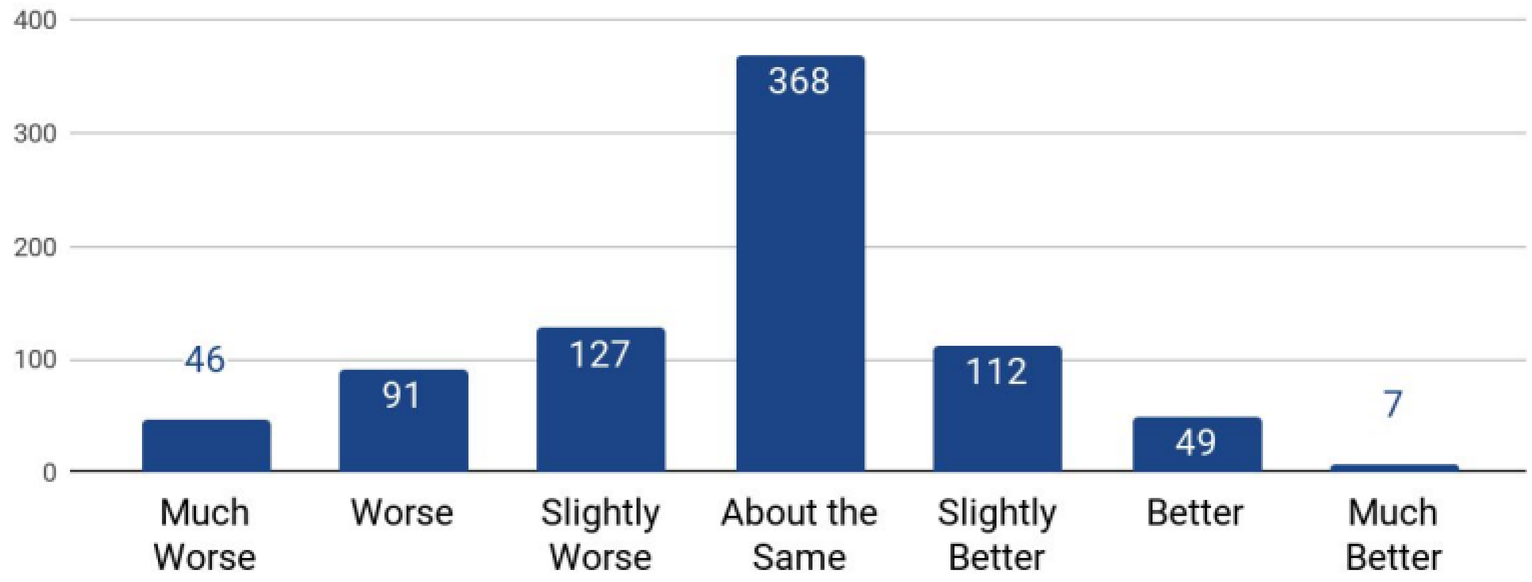


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.

Tactotron2 Examples

- https://google.github.io/tacotron/publications/tacotron2/index.html

2019

Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, Bhuvana Ramabhadran

Google

{ngyuzh, ronw}@google.com

Uses Tacotron

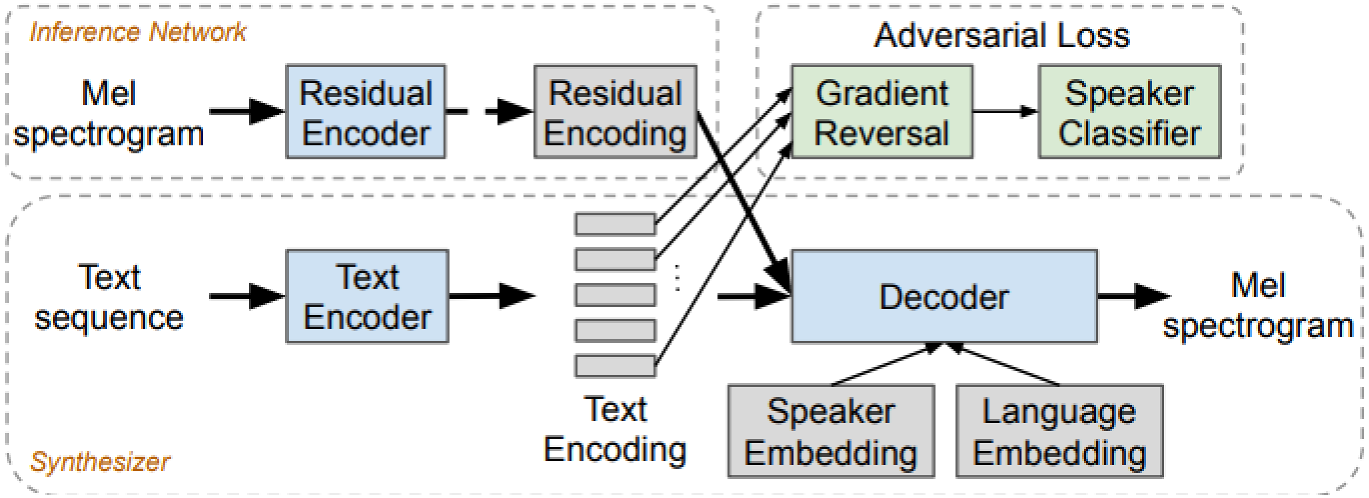


Figure 1: *Overview of the components of the proposed model. Dashed lines denote sampling via reparameterization [21] during training. The prior mean is always use during inference.*

Table 1: *Speaker similarity Mean Opinion Score (MOS) comparing ground truth audio from speakers of different languages. Raters are native speakers of the target language.*

Source Language	Target Language		
	EN	ES	CN
EN	**4.40±0.07**	1.72±0.15	1.80±0.08
ES	1.49±0.06	**4.39±0.06**	2.14±0.09
CN	1.32±0.06	2.06±0.09	**3.51±0.12**

Table 4: *Naturalness and speaker similarity MOS of cross-language voice cloning of the full multilingual model using phoneme inputs.*

Source Language	Model	EN target		ES target		CN target	
		Naturalness	Similarity	Naturalness	Similarity	Naturalness	Similarity
-	Ground truth (self-similarity)	4.60±0.05	4.40±0.07	4.37±0.06	4.39±0.06	4.42±0.06	3.51±0.12
EN	84EN 3ES 5CN	4.37±0.12	4.63±0.06	4.20±0.07	3.50±0.12	3.94±0.09	3.03±0.10
	language ID fixed to EN	-	-	3.68±0.07	4.06±0.09	3.09±0.09	3.20±0.09
ES	84EN 3ES 5CN	4.28±0.10	3.24±0.09	4.37±0.04	4.01±0.07	3.85±0.09	2.93±0.12
CN	84EN 3ES 5CN	4.49±0.08	2.46±0.10	4.56±0.08	2.48±0.09	4.09±0.10	3.45±0.12

2020

One Model, Many Languages: Meta-learning for Multilingual Text-to-Speech

Tomáš Nekvinda, Ondřej Dušek

Charles University, Faculty of Mathematics and Physics, Prague, Czechia
tom@neqindi.cz, odusek@ufal.mff.cuni.cz

Tacotron 2

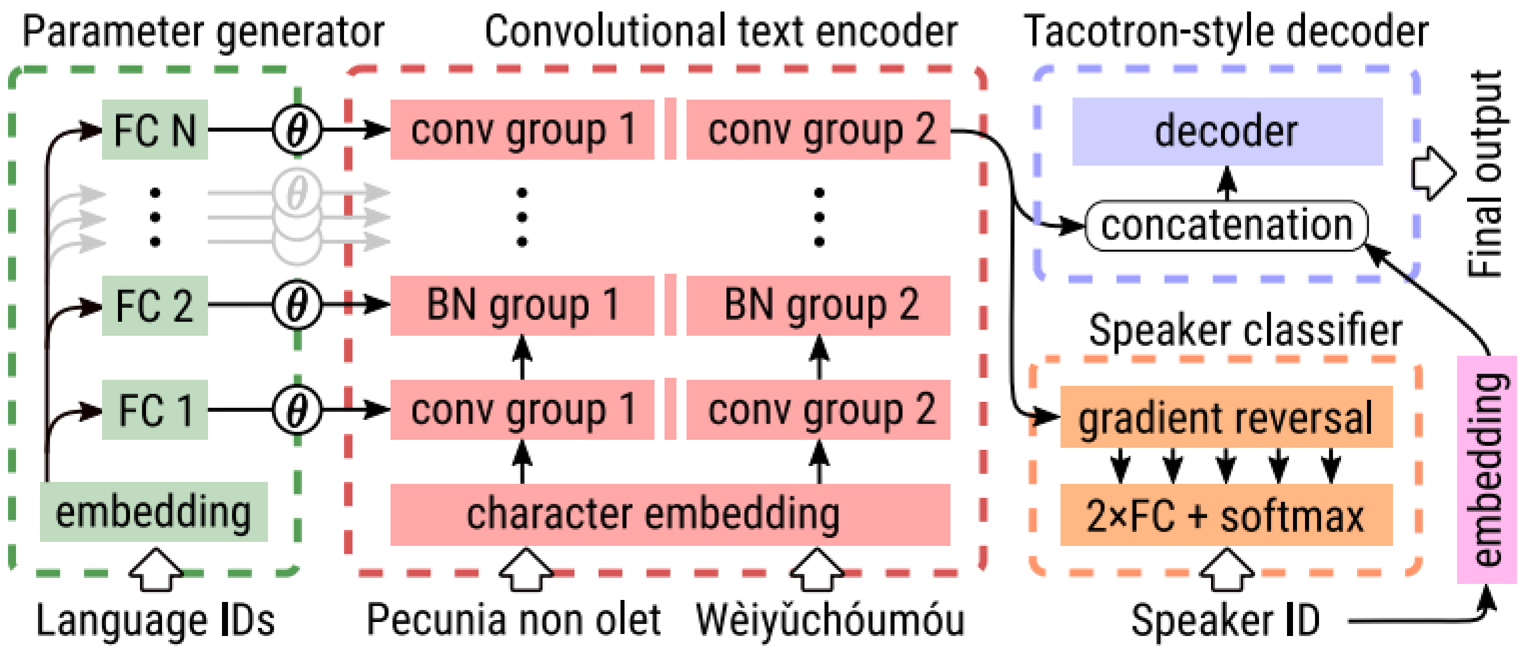


Figure 1: *Diagram of our model. The meta-network generates parameters of language-specific convolutional text encoders. Encoded text inputs enhanced with speaker embeddings are read by the decoder. The adversarial classifier suppresses speaker-dependent information in encoder outputs.*

For training, we used the CSS10 dataset and our new small dataset based on Common Voice recordings in five languages

Table 1: *Total data sizes per language (hours of audio data) in our cleaned CSS10 (CSS) and Common Voice (CV) subsets.*

	DE	EL	SP	FI	FR	HU	JP	NL	RU	ZH
CSS	15.4	3.5	20.9	9.7	16.9	9.5	14.3	11.7	17.7	5.6
CV	4.8	*N/A*	*N/A*	*N/A*	3.0	*N/A*	*N/A*	1.3	3.4	1.0

Table 2: *Left: CERs of ground-truth recordings (GT) and recordings produced by monolingual and the three examined multilingual models. Right: CERs of the recordings synthesized by* GEN *and* SHA *trained on just 600 or 900 training examples per language. Best results for the given language are shown in bold; "*" denotes statistical significance (established using paired t-test; $p < 0.05$).*

	GT	SGL	SHA	SEP	GEN	SHA 600	SHA 900	GEN 600	GEN 900
DE	4.8 ± 4.6	7.3 ± 6.0	8.3 ± 6.0	15.3 ± 6.0	***5.8 ± 5.3**	13.2 ± 8.9	**12.4 ± 8.0**	15.6 ± 9.4	12.5 ± 9.3
EL	8.7 ± 6.9	*N/A*	**11.4 ± 8.3**	22.2 ± 8.3	11.6 ± 7.1	16.8 ± 9.7	16.0 ± 10.2	**14.2 ± 8.7**	14.7 ± 9.8
SP	3.9 ± 4.6	7.0 ± 10.8	7.2 ± 6.5	10.2 ± 8.1	**7.0 ± 9.8**	9.8 ± 7.5	9.9 ± 8.4	8.1 ± 6.0	***7.6 ± 5.9**
FI	6.9 ± 10.4	18.6 ± 12.6	**10.3 ± 8.0**	18.1 ± 11.4	10.4 ± 7.0	18.2 ± 12.2	18.4 ± 13.2	***13.2 ± 10.9**	14.0 ± 10.6
FR	11.2 ± 7.3	25.2 ± 12.6	30.0 ± 14.3	54.5 ± 21.9	***19.0 ± 12.9**	40.2 ± 15.8	37.6 ± 16.2	32.9 ± 13.2	***27.2 ± 12.2**
HU	6.3 ± 6.1	15.8 ± 9.5	15.9 ± 10.6	18.8 ± 9.9	***13.5 ± 8.3**	21.4 ± 10.4	21.3 ± 13.0	***16.5 ± 10.4**	18.0 ± 10.4
JP	19.0 ± 9.3	28.8 ± 11.3	27.2 ± 11.8	33.7 ± 13.5	**25.1 ± 12.2**	32.5 ± 12.8	32.2 ± 15.0	**29.9 ± 13.0**	30.9 ± 13.5
NL	14.5 ± 7.4	33.4 ± 13.8	31.6 ± 12.5	49.0 ± 17.4	***22.6 ± 9.6**	37.8 ± 13.5	30.4 ± 10.2	32.8 ± 12.3	**28.3 ± 9.8**
RU	12.3 ± 15.0	45.5 ± 24.1	44.4 ± 21.9	58.1 ± 24.7	***34.5 ± 21.3**	60.4 ± 18.6	47.0 ± 20.5	38.5 ± 20.1	***34.4 ± 17.9**
ZH	14.6 ± 11.8	62.8 ± 18.5	28.6 ± 15.9	27.3 ± 14.8	***20.5 ± 13.6**	40.2 ± 15.2	39.8 ± 18.8	33.0 ± 15.5	***28.4 ± 15.6**

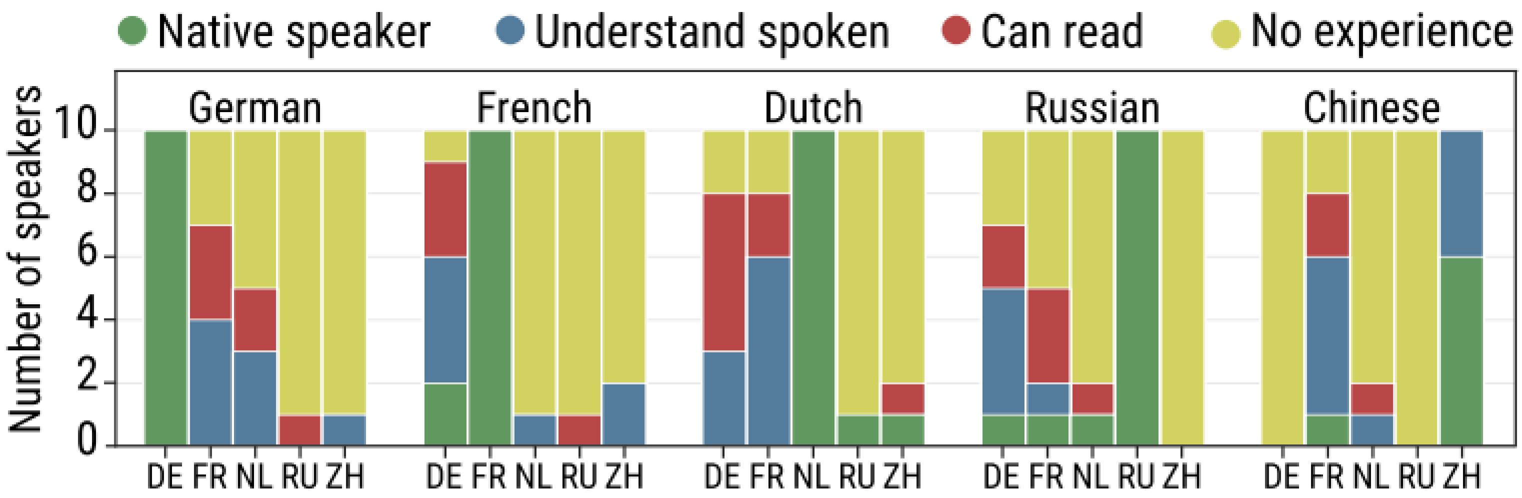


Figure 2: *Language abilities of participants of our survey.*

Table 3: *Mean (with std. dev.) ratings of fluency, naturalness, voice stability (top) and pronunciation accuracy (middle). The bottom row shows the number of sentences with word skips.*

		SHA	SEP	GEN
Fluency	German	3.0 ± 1.1	2.6 ± 1.0	*$\mathbf{3.4 \pm 0.9}$
	French	2.8 ± 1.0	2.6 ± 1.0	*$\mathbf{3.5 \pm 0.9}$
	Dutch	3.1 ± 0.9	2.5 ± 1.1	*$\mathbf{3.7 \pm 1.0}$
	Russian	2.8 ± 1.0	2.5 ± 1.0	*$\mathbf{3.4 \pm 0.9}$
	Chinese	2.7 ± 1.3	2.6 ± 1.2	*$\mathbf{3.5 \pm 1.2}$
	All	2.9 ± 1.1	2.5 ± 1.1	*$\mathbf{3.5 \pm 1.0}$
Accuracy	German	3.3 ± 1.1	3.1 ± 1.2	*$\mathbf{3.7 \pm 1.0}$
	French	3.1 ± 1.1	2.7 ± 1.2	*$\mathbf{3.7 \pm 0.9}$
	Dutch	3.4 ± 1.0	2.5 ± 1.2	*$\mathbf{3.9 \pm 1.1}$
	Russian	3.0 ± 1.2	2.6 ± 1.2	*$\mathbf{3.6 \pm 1.0}$
	Chinese	2.9 ± 1.4	2.8 ± 1.4	*$\mathbf{3.5 \pm 1.2}$
	All	3.1 ± 1.2	2.7 ± 1.2	*$\mathbf{3.7 \pm 1.1}$
Word skips		41/400	38/400	**11/400**

Code-switching evaluation dataset: We created a new small-scale dataset especially for code-switching evaluation. We used bilingual sentences scraped from Wikipedia. For each language, we picked 80 sentences with a few foreign words (20 sentences for each of the 4 other languages); Chinese was romanized. We replaced foreign names with their native forms (see Fig. 3).

● German ● Russian ● Dutch ● French ● Chinese

Der кремль ist das wichtigste Bauwerk in der Нижний Новгород Altstadt.

gànzhōushì est une ville du sud de la province du jiāngxīshěng en Chine.

De Oberbürgermeister slaat onder veel belangstelling het eerste vat bier aan

Figure 3: *Examples of code-switching evaluation sentences.*

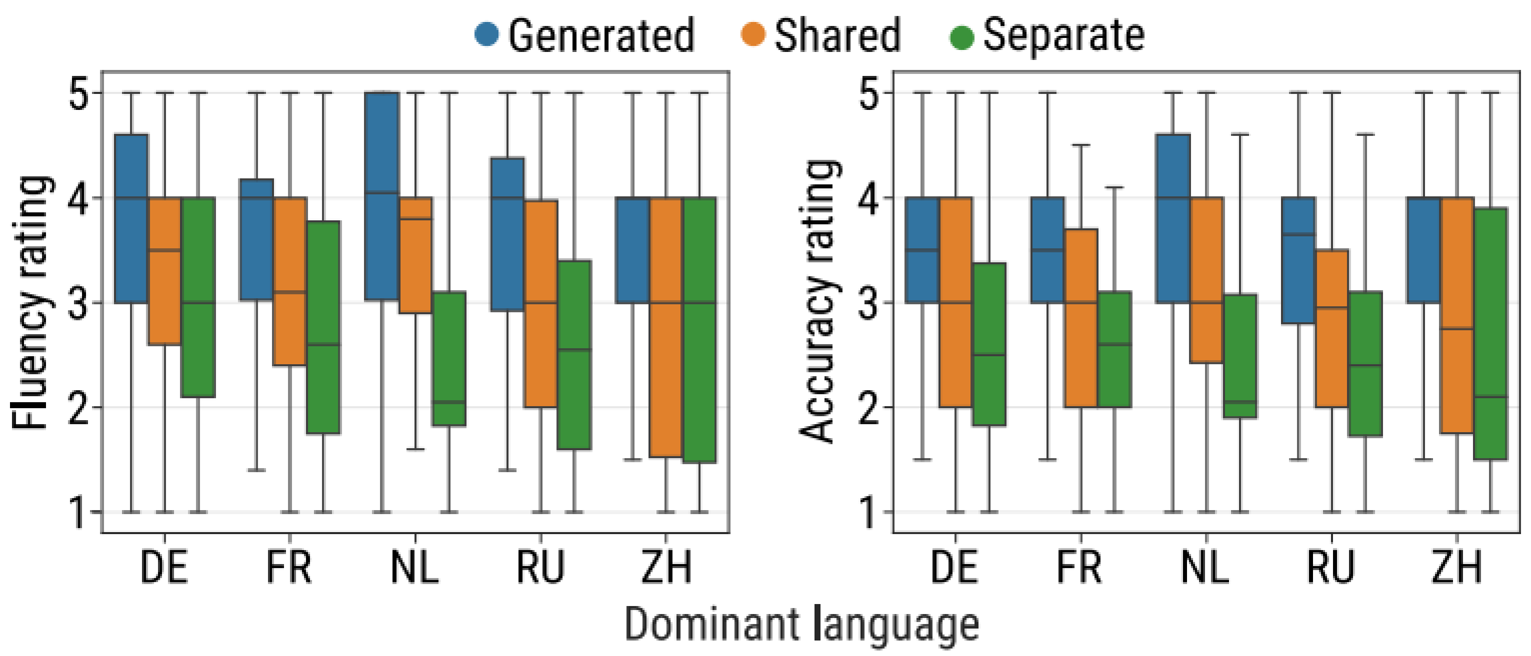


Figure 4: *Graphs showing distributions of fluency and accuracy ratings grouped by the dominant language of rated sentences.*

2020

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Towards Universal Text-to-Speech

Jingzhou Yang and Lei He

Microsoft, China
{jingy,helei}@microsoft.com

- 1,250 hours of data from 50 language locales
- Data in different locales is highly unbalanced → Balance
- 20 seconds of data is feasible for a new speaker
- 6 minutes for a new language

“The neural vocoder can be any vocoder that converts mel spectrograms to waveforms, e.g. WaveNet [20], WaveRNN [21] or LPCNet [22]. WaveNet is used in this paper.”

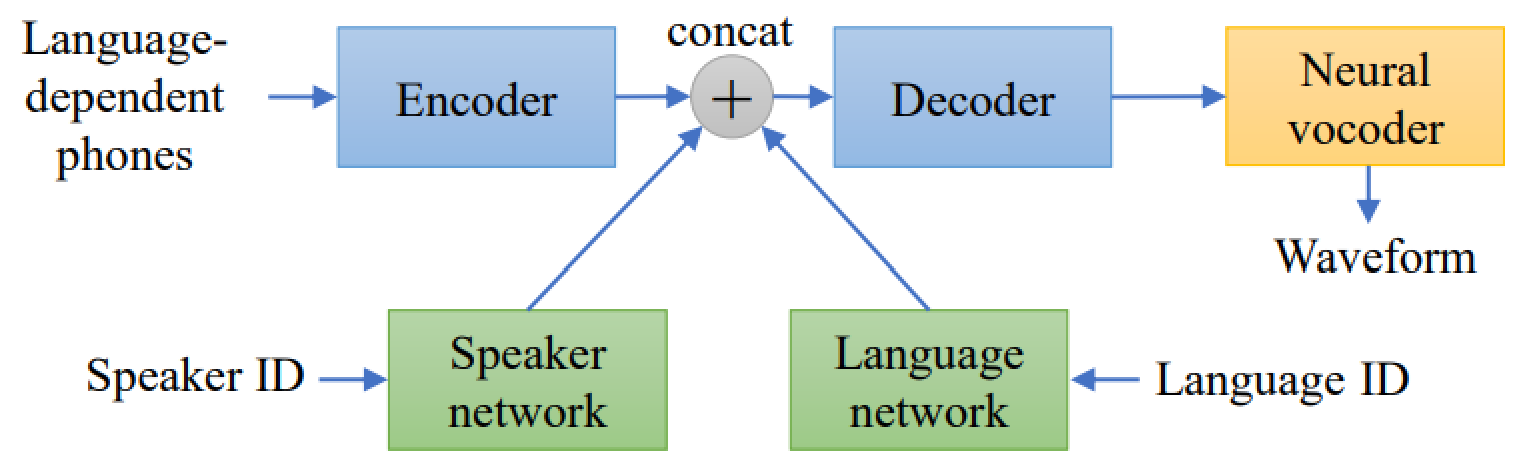


Figure 1: *The framework of the multilingual system.*

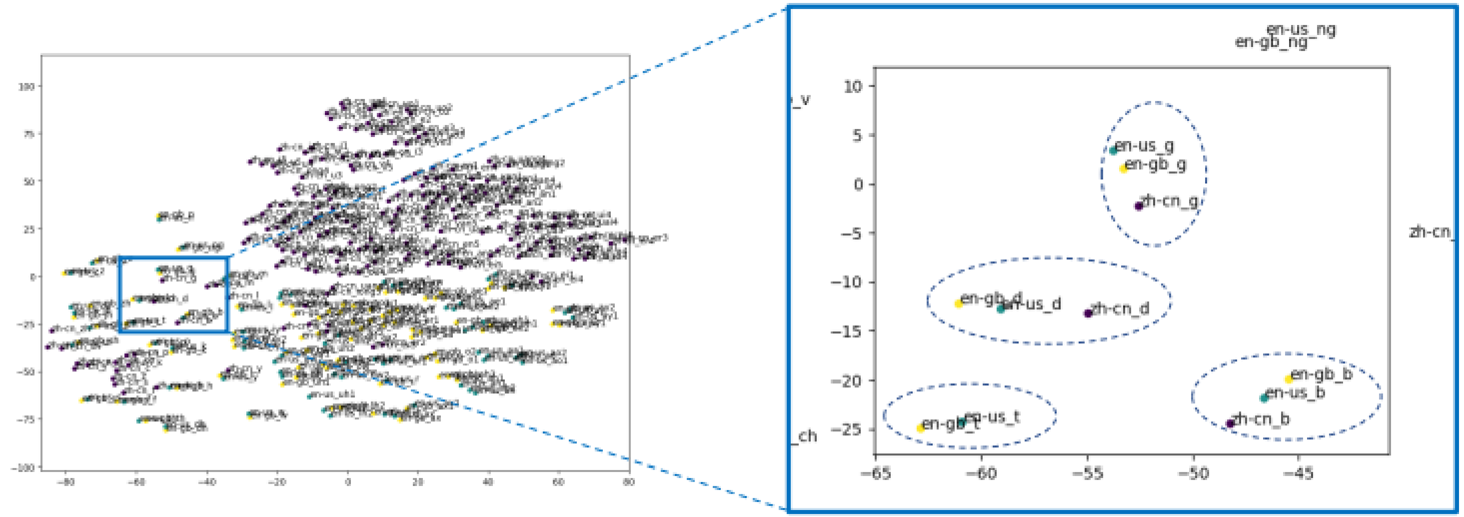


Figure 2: *The t-SNE visualization of the phone embeddings.*

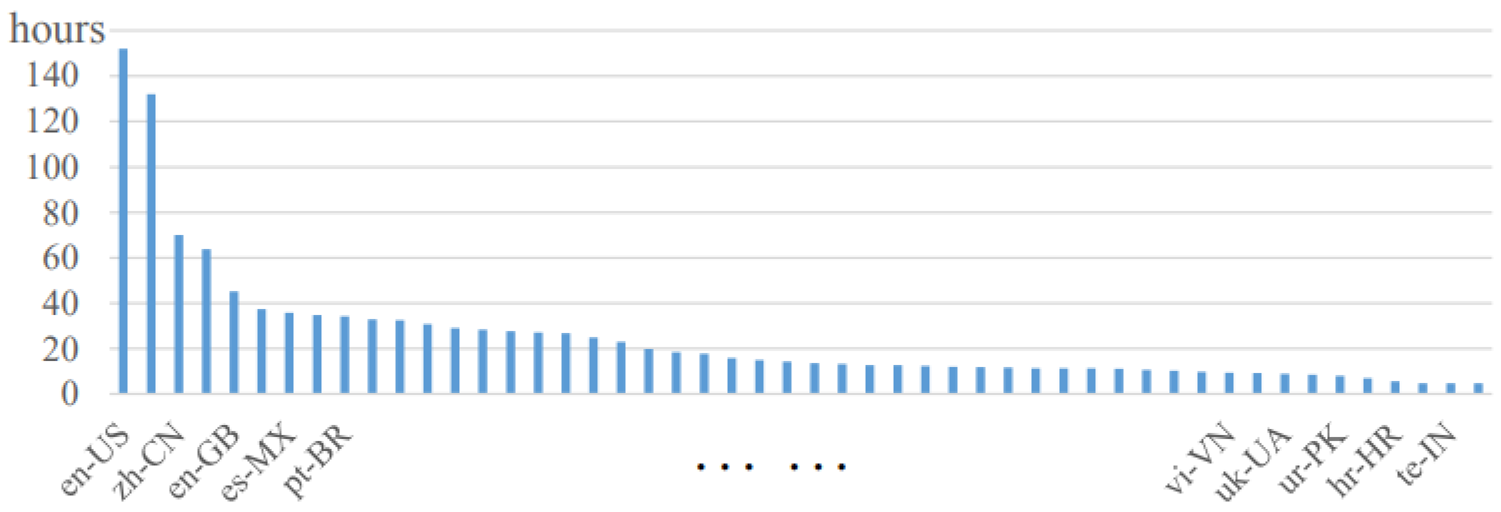


Figure 3: *The data distribution over 50 language locales.*

Table 1: *The naturalness MOS in different languages.*

Language Data size	en-US 20h/150h	de-DE 10h/30h	vi-VN 7h/7h	te-IN 5h/5h
Rec.	4.51±0.10	4.22±0.13	4.23±0.15	4.47±0.13
Single	4.34±0.08	4.19±0.08	4.14±0.09	3.40±0.13
Multi	**4.30±0.08**	4.07±0.08	3.83±0.10	3.59±0.12
+LgB	4.03±0.09	**4.08±0.08**	**4.03±0.09**	**3.89±0.11**
+SpkB	4.19±0.08	4.03±0.09	3.90±0.09	3.73±0.11

Table 2: *The naturalness MOS to the de-DE speaker.*

Language	en-US	vi-VN	te-IN
Rec.	4.55±0.09*	4.50±0.11*	4.59±0.14*
Multi	**3.97±0.10**	3.78±0.09	3.54±0.13
+LgB	3.86±0.09	**3.79±0.07**	**3.79±0.11**

Table 3: *The similarity MOS to the de-DE speaker.*

Language	en-US	vi-VN	te-IN
Rec.	1.27±0.08*	1.12±0.07*	1.52±0.12*
Multi	2.93±0.19	2.69±0.17	2.70±0.17
+LgB	2.98±0.19	2.50±0.18	2.47±0.16

Table 4: *The MOS to the new zh-CN speaker.*

	Naturalness		Similarity	
Language	zh-CN	en-US	zh-CN	en-US
Rec.	3.78±0.13	3.37±0.20	4.32±0.12	3.77±0.12
20s	3.61±0.07	3.72±0.08	4.21±0.12	3.43±0.12
1m	3.62±0.07	3.76±0.08	4.32±0.10	3.49±0.11
5m	3.68±0.07	3.71±0.08	4.20±0.12	3.35±0.12
10m	3.63±0.07	3.61±0.09	4.27±0.11	3.25±0.14

Table 5: *The MOS to the new en-GB speaker.*

	Naturalness		Similarity	
Language	en-GB	zh-CN	en-GB	zh-CN
Rec.	4.56±0.11	–	4.49±0.11	–
20s	4.08±0.08	3.61±0.07	4.36±0.12	2.60±0.24
1m	4.16±0.09	3.57±0.08	4.42±0.12	2.26±0.23
5m	4.24±0.08	3.34±0.07	4.47±0.12	2.30±0.23
10m	4.24±0.08	3.19±0.08	4.36±0.13	2.36±0.23