Attention-Based Models for Speech Recognition Supplementary Material

Jan Chorowski
University of Wrocław, Poland
jan.chorowski@ii.uni.wroc.pl

Dzmitry Bahdanau
Jacobs University Bremen, Germany

Dmitriy Serdyuk
Université de Montréal

Kyunghyun Cho
Université de Montréal

Yoshua Bengio
Université de Montréal
CIFAR Senior Fellow

1 Additional Figures

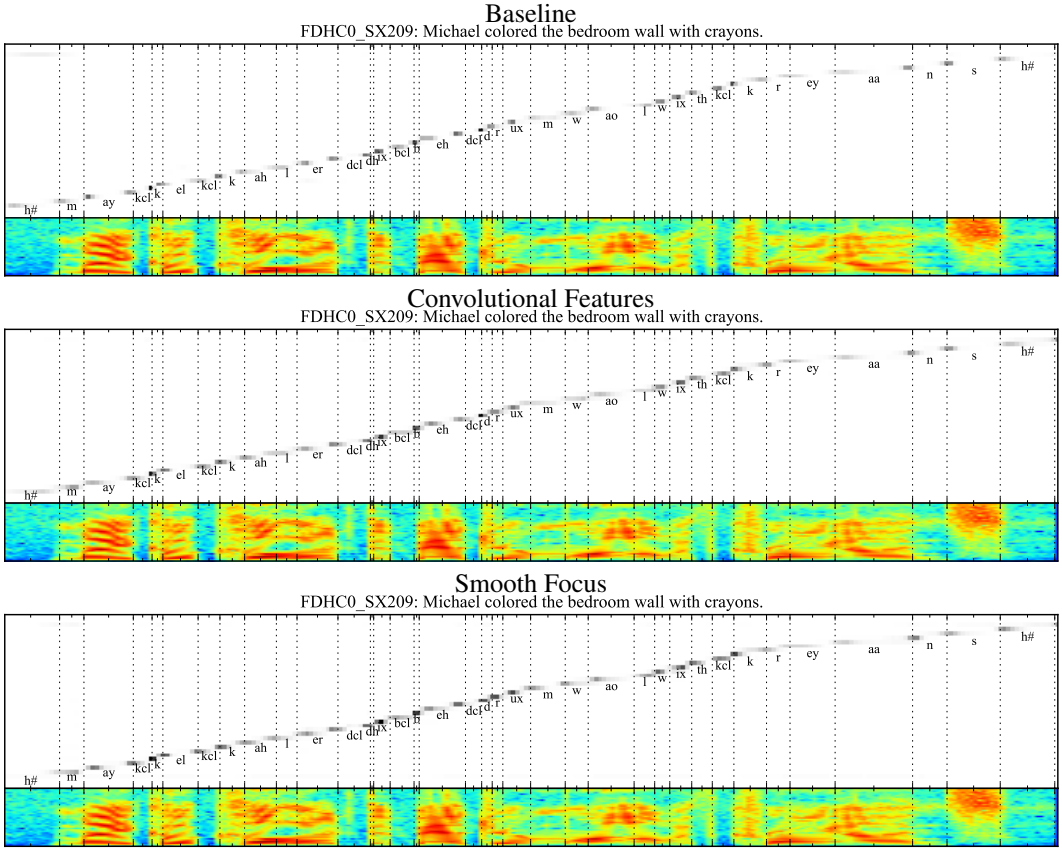

Figure 1: Alignments produced by evaluated models on the FDHC0_SX209 test utterance. The vertical bars indicate ground truth phone location from TIMIT. Each row of the upper image indicates frames selected by the attention mechanism to emit a phone symbol. Compare with Figure 3. in the main text.

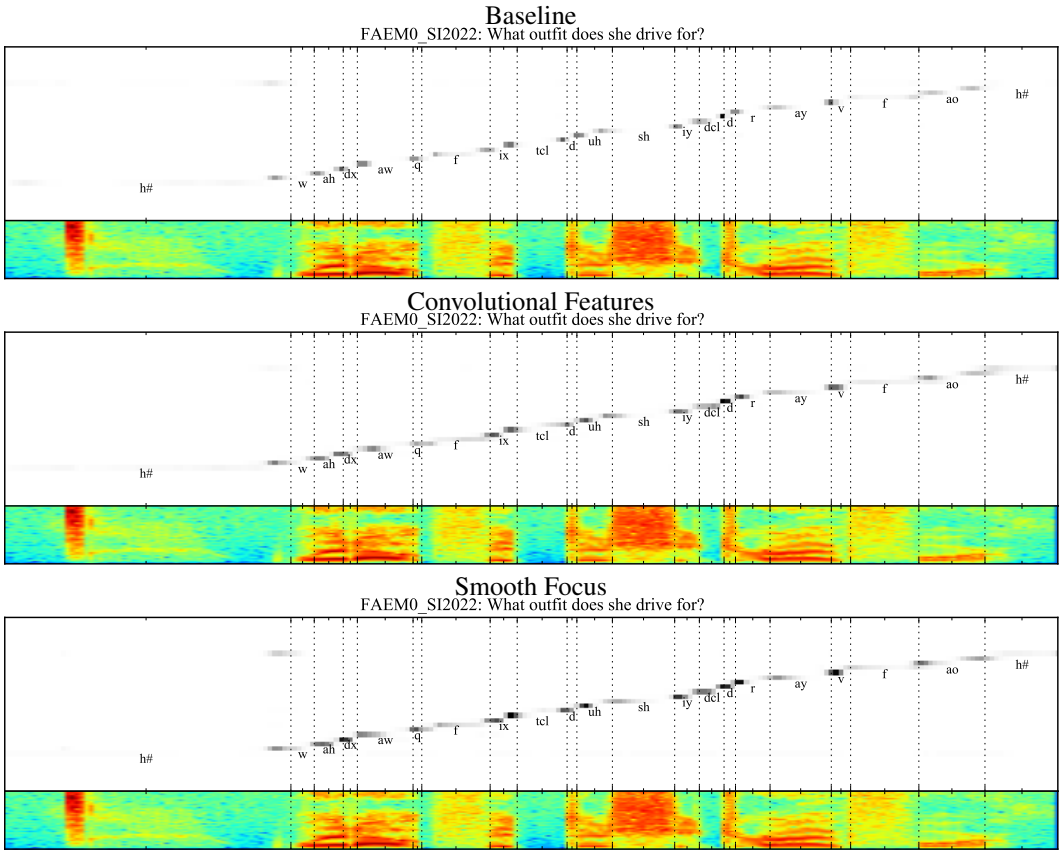

Figure 2: Alignments produced by evaluated models on the FAEM0_SI2022 train utterance. The vertical bars indicate ground truth phone location from TIMIT. Each row of the upper image indicates frames selected by the attention mechanism to emit a phone symbol. Compare with Figure 3. in the main text.

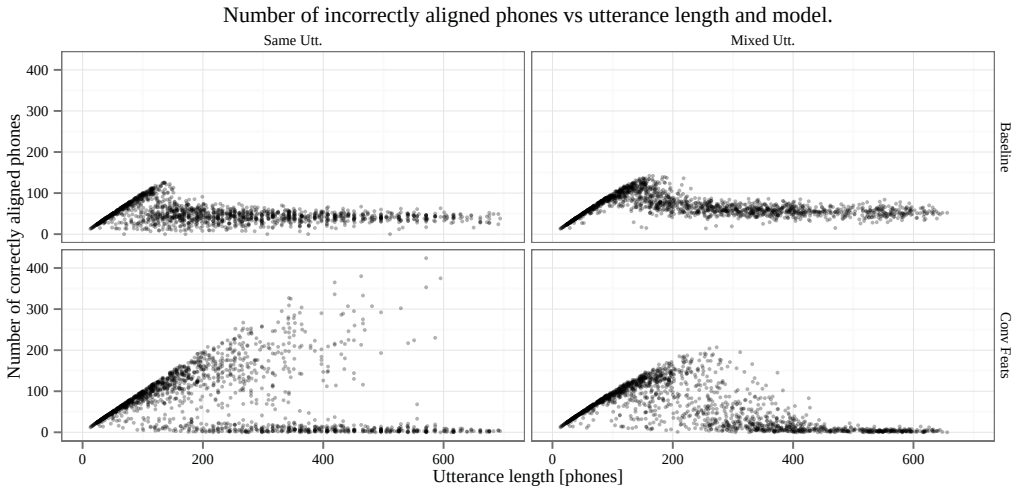

Figure 3: Close-up on the two failure modes of ARSG. Results of force-aligning concatenated TIMIT utterances. Each dot represents a single utterance. The left panels show results for concatenations of the same utterance. The right panels show results for concatenations of randomly chosen utterances. We compare the baseline network having a content-based only attention mechanism (top row) with a hybrid attention mechanism that uses convolutional features (bottom row). While neither model is able to properly align long sequences, they fail in different ways: the baseline network always aligns about 50 phones, while the location-aware network fails to align any phone. Compare with Figure 4 form the main paper.

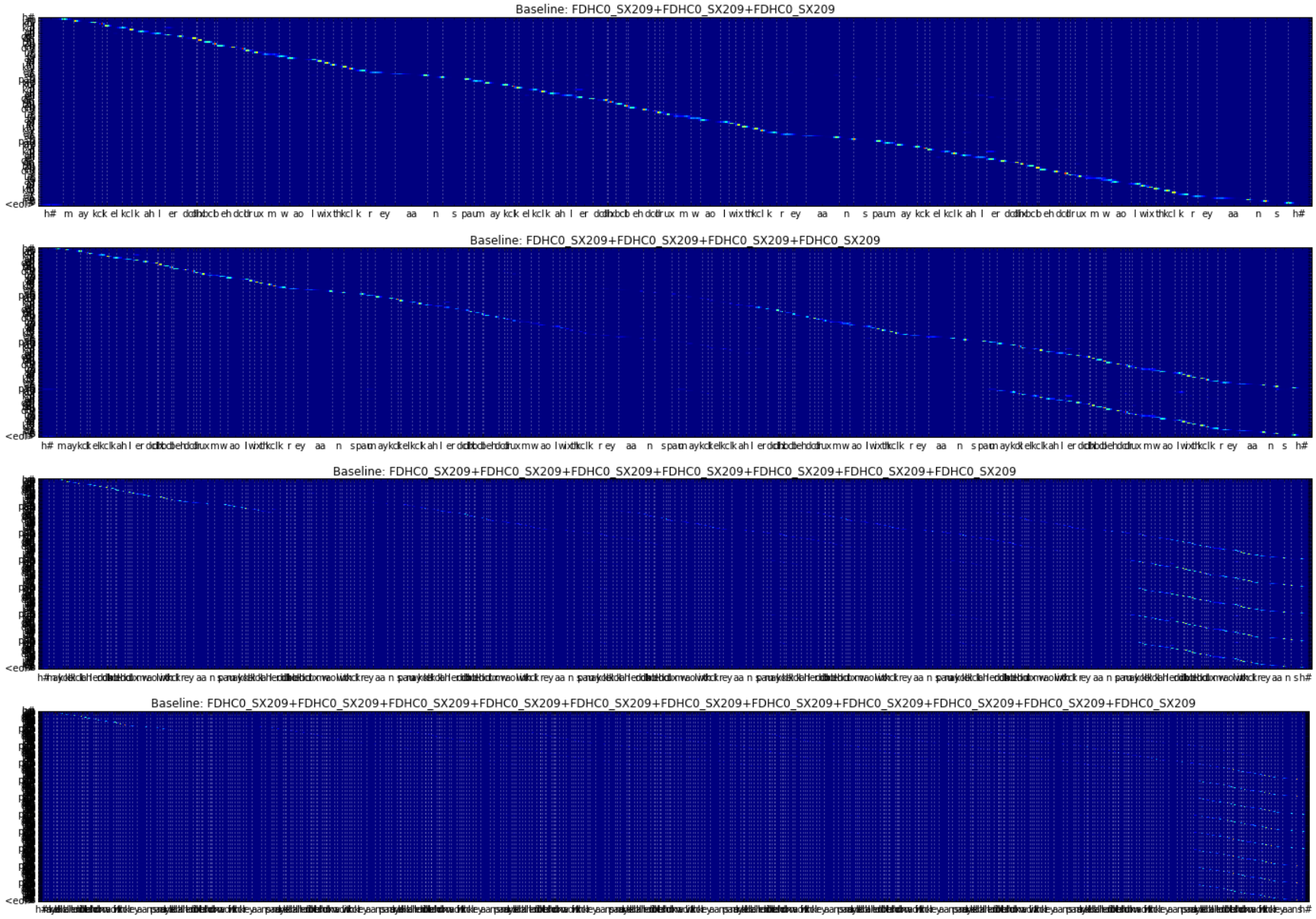

Figure 4: The baseline network fails to align more than 3 repetitions of FDHC0_SX209.

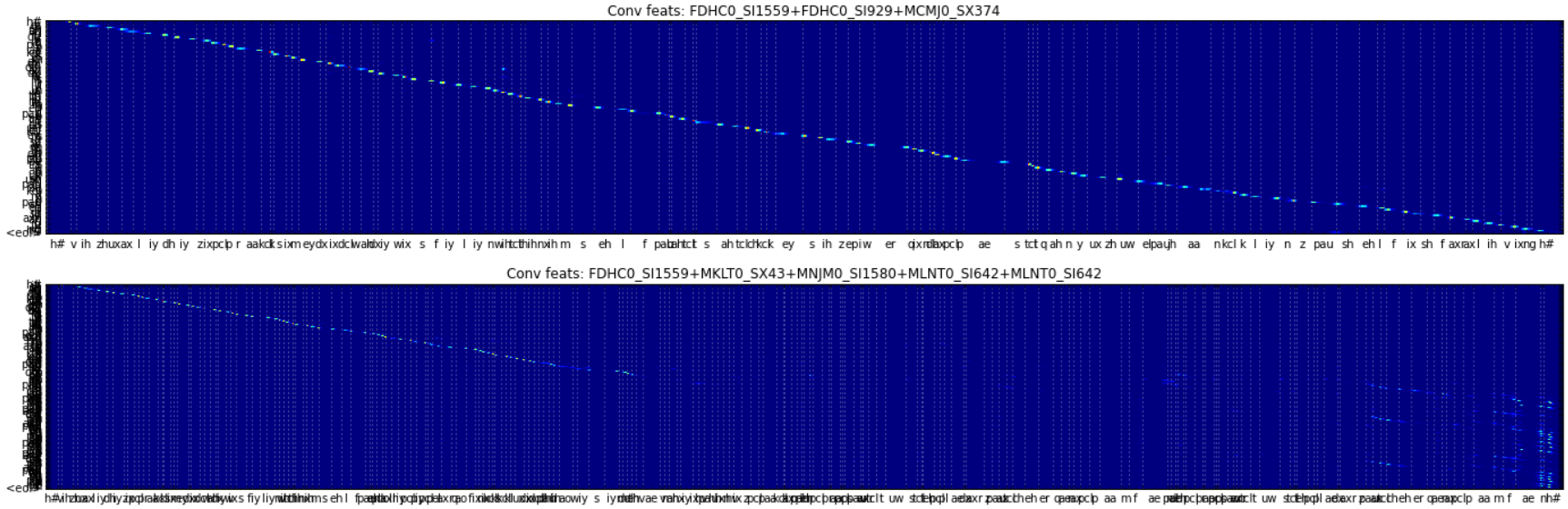

Figure 5: The baseline network aligns a concatenation of 3 different utterances, but fails to align 5.

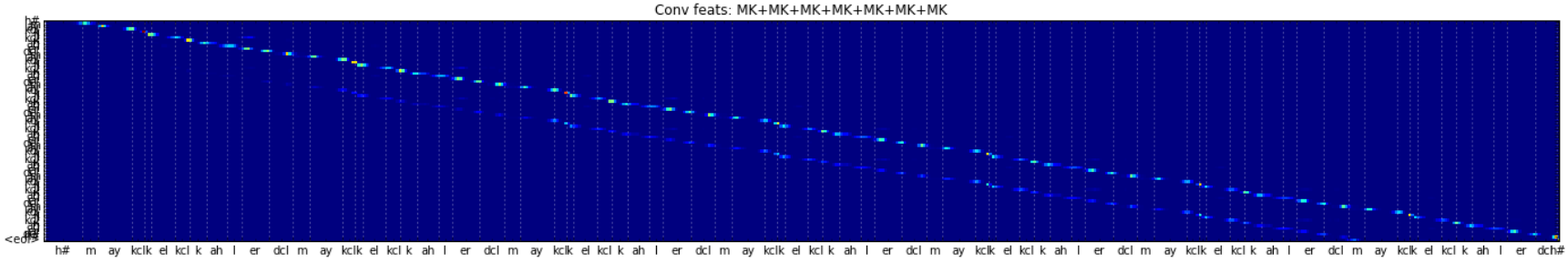

Figure 6: Forced alignment of 7 repetitions of the phrase "Michael colored" performed with the baseline model with windowing enabled (the alignment was constrained to ± 75 frames from the expected position of the generator at the last step. The window is wider than the pattern and the net confuses similar content. Strangely, the first two repetitions are aligned without any confusion with subsequent ones – the network starts to confound phoneme location only starting from the third repetition (as seen by the parallel strand of alignment which starts when the network starts to emit the phrase for the third time).

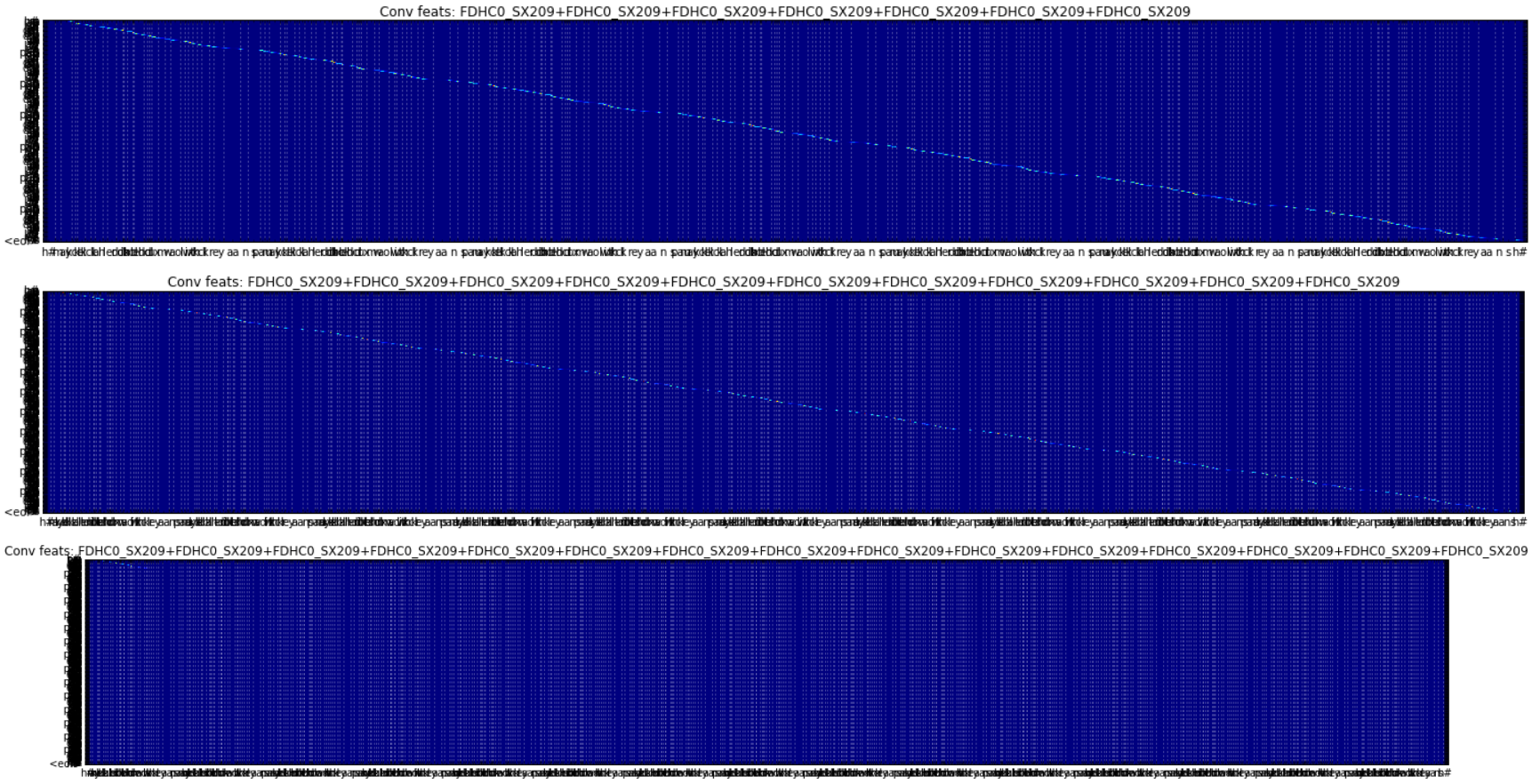

Figure 7: The location-aware network correctly aligns 7 and 11 repetitions of FDHC0_SX209, butfails to align 15 repetitions of FDHC0_SX209.

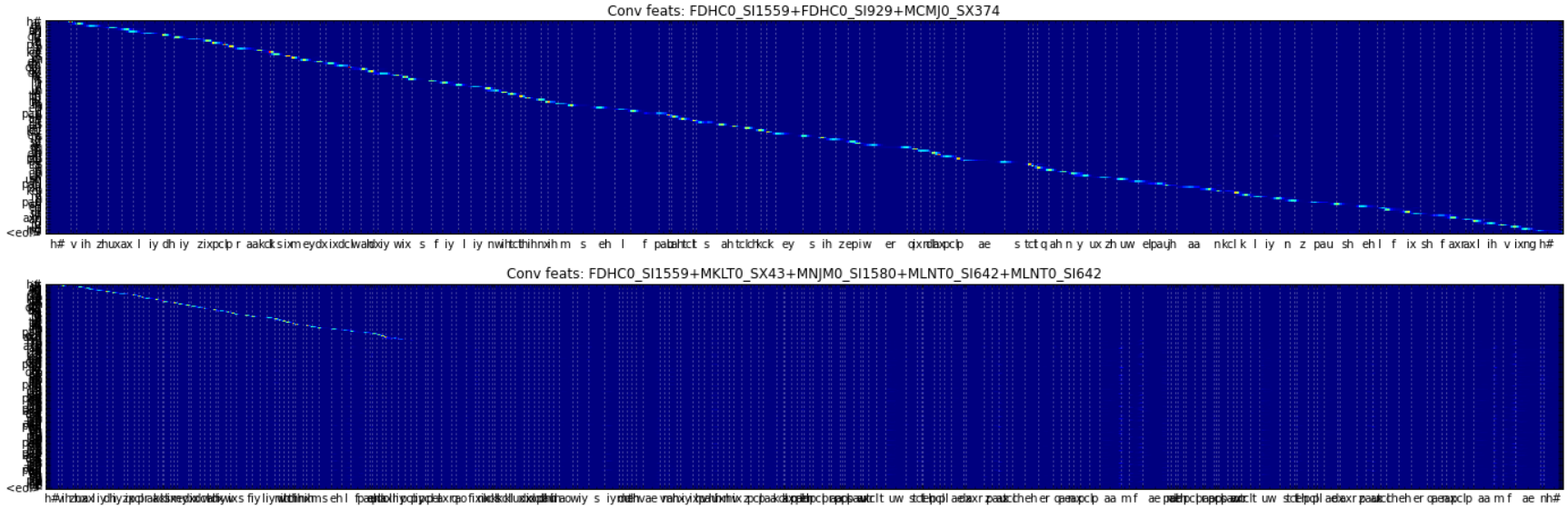

Figure 8: The location-aware network aligns a concatenation of 3 different utterances, but fails to align 5.

2 Detailed results of experiments

Table 1: Phoneme error rates while decoding with various modifications. Compare with Figure 5 from the main paper.

		Plain	Keep 1	Keep 10	Keep 50	$\beta = 2$	Win. ± 75	Win. ± 150
Baseline	dev	15.9%	17.6%	15.9%	15.9%	16.1%	15.9%	15.9%
	test	18.7%	20.2%	18.7%	18.7%	18.9%	18.7%	18.6%
Conv Feats	dev	16.1%	19.4%	16.2%	16.1%	16.7%	16.0%	16.1%
	test	18.0%	22.3%	17.9%	18.0%	18.7%	18.0%	18.0%
Smooth Focus	dev	15.8%	21.6%	16.5%	16.1%	16.2%	16.2%	16.0%
	test	17.6%	24.7%	18.7%	17.8%	18.4%	17.7%	17.6%