STC innovations

The 5th International Workshop on Speech Processing in Everyday Environments

THE STC SYSTEM FOR THE CHIME 2018 CHALLENGE

Ivan Medennikov, Ivan Sorokin, Aleksei Romanenko, Dmitry Popov, Yuri Khokhlov, Tatiana Prisyach, Nikolay Malkovskii, Vladimir Bataev, Sergei Astapov, Maxim Korenevsky, Alexander Zatvornitskiy

WHO WE ARE

STC-INNOVATIONS

1. Top-3 in Babel OpenKWS, 1st NIST i-vector Machine Learning Challenge 2014, 2nd NIST LRE 2015, 2nd NIST SITW, 2015 2nd ANTISPOOF 2015, 2017 1nd ANTISPOOF 2017
2. Multi-disciplinary team with expertise in general machine learning, speech recognition, NLU, bi-modal (voice+face) identification
3. Close partnership with ITMO University

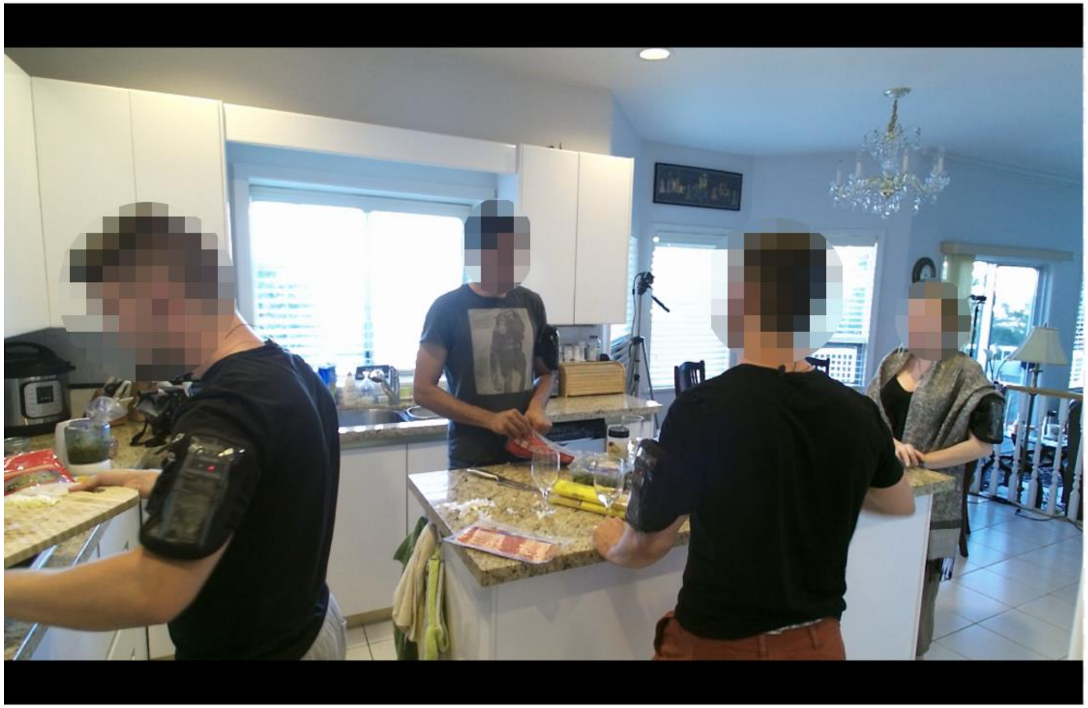
Outline

- Introduction
- Unsuccess story
- Success story
- Conclusions
- Final results on eval and future work

Introduction

Main challenges

- Conversational speech
- Noisy real-world environment
- Far-field conditions
- Great amount of overlapped speech



Beamforming and Enhancement: Unsuccess story

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] *

*https://github.com/auspicious3000/deepbeam

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016]*

*https://github.com/fgnt/nn-gev

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗

STC innovations

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017]

STC innovations

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017] ✗

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017] ✗
- Deep Clustering

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017] ✗
- Deep Clustering ✗

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017] ✗
- Deep Clustering ✗
- Permutation invariant training (PIT)

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017] ✗
- Deep Clustering ✗
- Permutation invariant training (PIT) ✗

Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017] ✗
- Deep Clustering ✗
- Permutation invariant training (PIT) ✗
- WPE

STC innovations

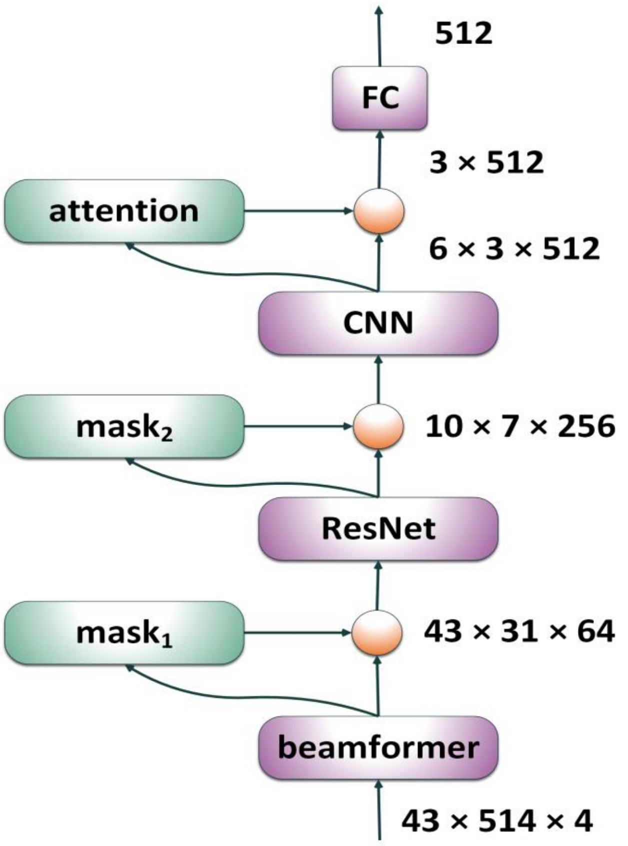
Unsuccess story

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017] ✗
- Deep Clustering ✗
- Permutation invariant training (PIT) ✗
- WPE ✓

Success story

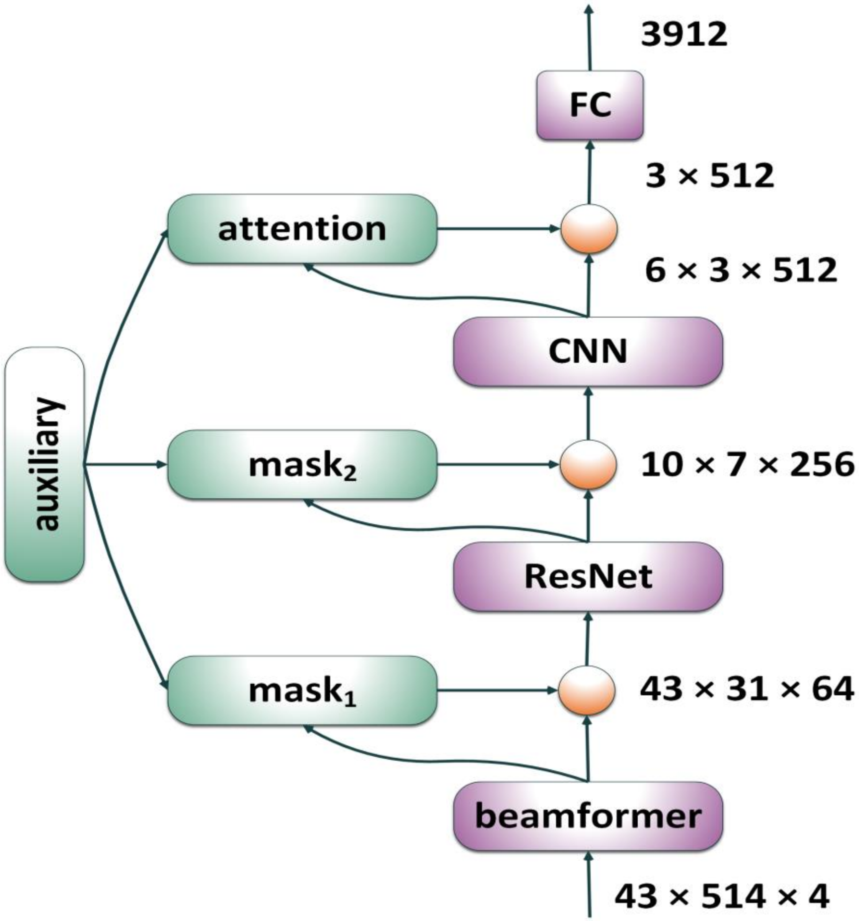
Multi-channel speaker-aware model training: embeddings

- embedding training by triplet ranking loss [Ye and Guo, 2018]



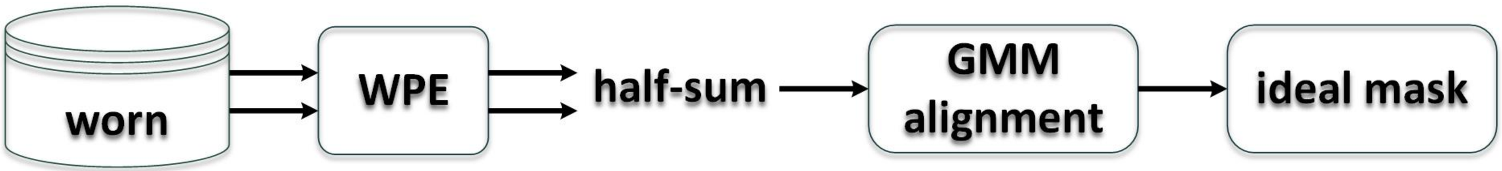
Multi-channel speaker-aware model training: final model

- auxiliary inputs [Zmolikova, 2018]
- residual attention network [Wang, 2017]
- speaker-adapted classifier *
- sum and average all embeddings for speaker in utterance



*https://github.com/Microsoft/LightGBM

Speaker adaptation by frame-level mask: training



	<sil>						word						<sil>				<noise>			<sil>				
P01(id 1)	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0
	<sil>		word			<sil>			<laught>						<sil>						word		<sil>	
P02(id 2)	0	0	2	2	2	0	0	0	2	2	2	2	2	2	0	0	0	0	0	0	2	2	0	0
	<sil>			word				<sil>		word			<sil>					<spn>			<sil>			
P03(id 4)	0	0	0	4	4	4	4	0	0	4	4	4	0	0	0	0	0	4	4	4	0	0	0	0
	<sil>								<laught>			<sil>			word			<sil>				word		
P04(id 8)	0	0	0	0	0	0	0	0	8	8	8	0	0	0	8	8	8	0	0	0	0	8	8	8
Ideal mask (general)	**0**	**0**	**2**	**6**	**6**	**4**	**5**	**1**	**11**	**15**	**15**	**7**	**2**	**2**	**8**	**8**	**9**	**5**	**5**	**4**	**2**	**10**	**8**	**8**
Ideal targets (if P01)	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**

Speaker adaptation by frame-level mask: filtering

Original acoustic feats	$x_{t,1}$	$x_{t+1,1}$	…					$x_{t+23,1}$
	⋮	⋮	⋱					⋮
	$x_{t,n}$	$x_{t+1,n}$	…					$x_{t+23,n}$
Speaker mask	**0.6**	**0.7**	**0.5**	**0.1**	**0.2**	**0.3**	**0.4**	**0.4**
Filtered acoustic feats	$x_{t,1}$	$x_{t+1,1}$	$x_{t+2,1}$	Throw out			$x_{t+22,1}$	$x_{t+23,1}$
	⋮	⋮	⋮				⋮	⋮
	$x_{t,n}$	$x_{t+1,n}$	$x_{t+2,n}$				$x_{t+22,n}$	$x_{t+23,n}$

* https://github.com/speechpro/mixup (for Kaldi)

Mixup [Medennikov, 2018] *

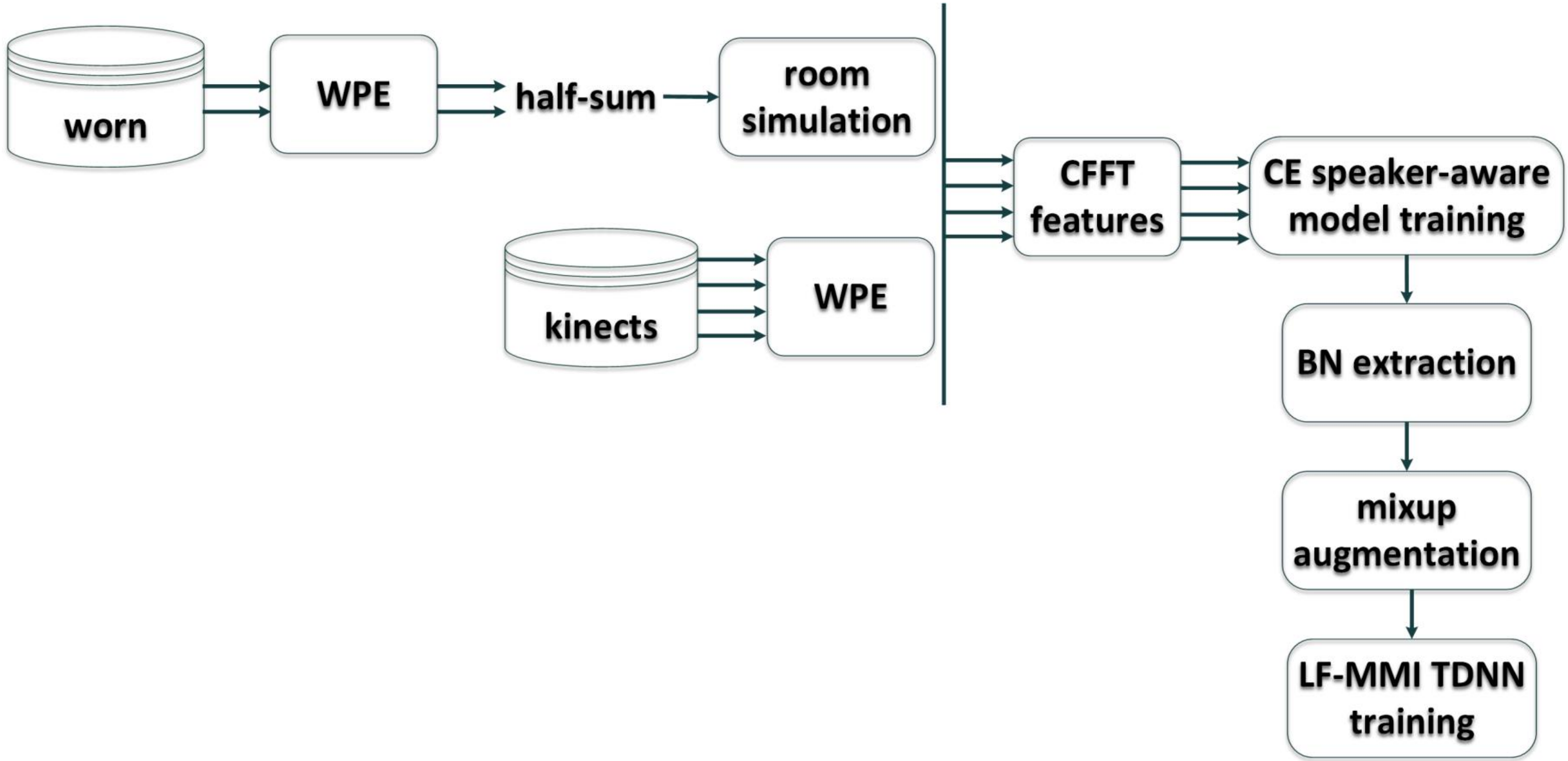
- virtual training examples by combining existing ones
- especially effective on mismatched test data

Generation of new training data

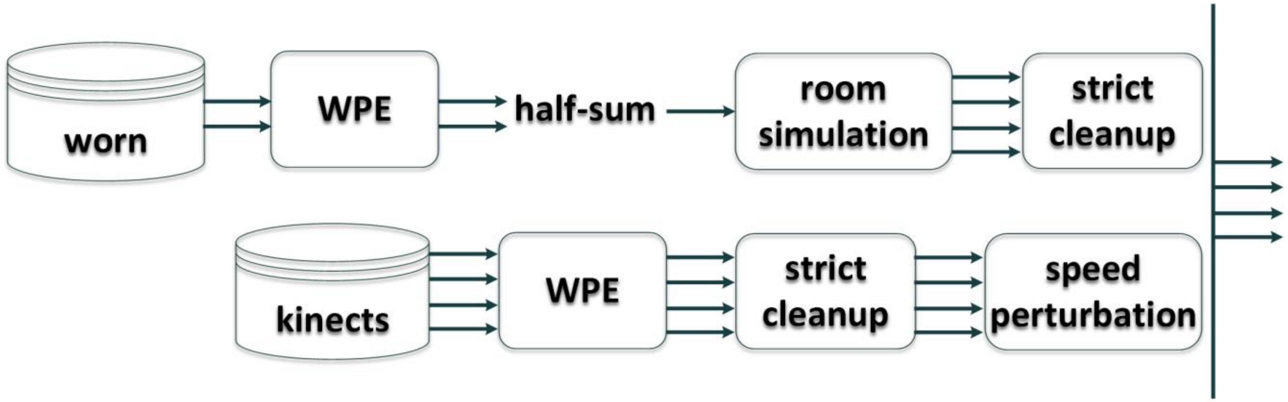
$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

Success story

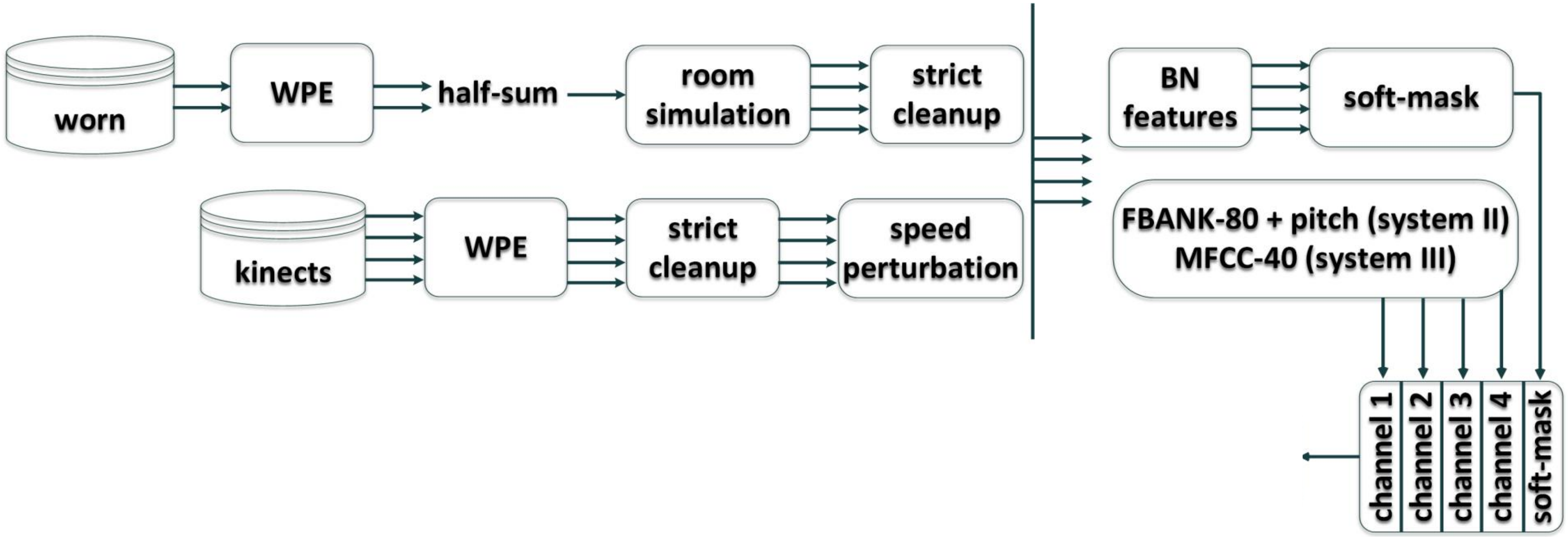
System I



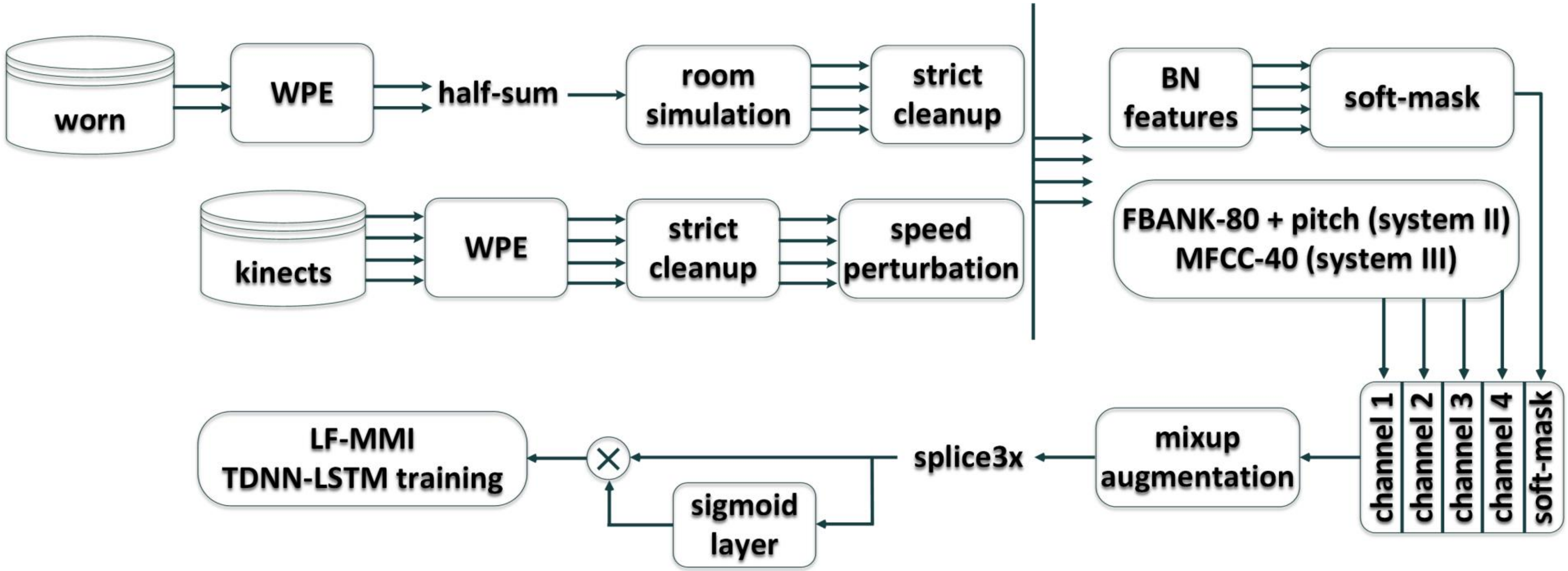
System II and III



System II and III

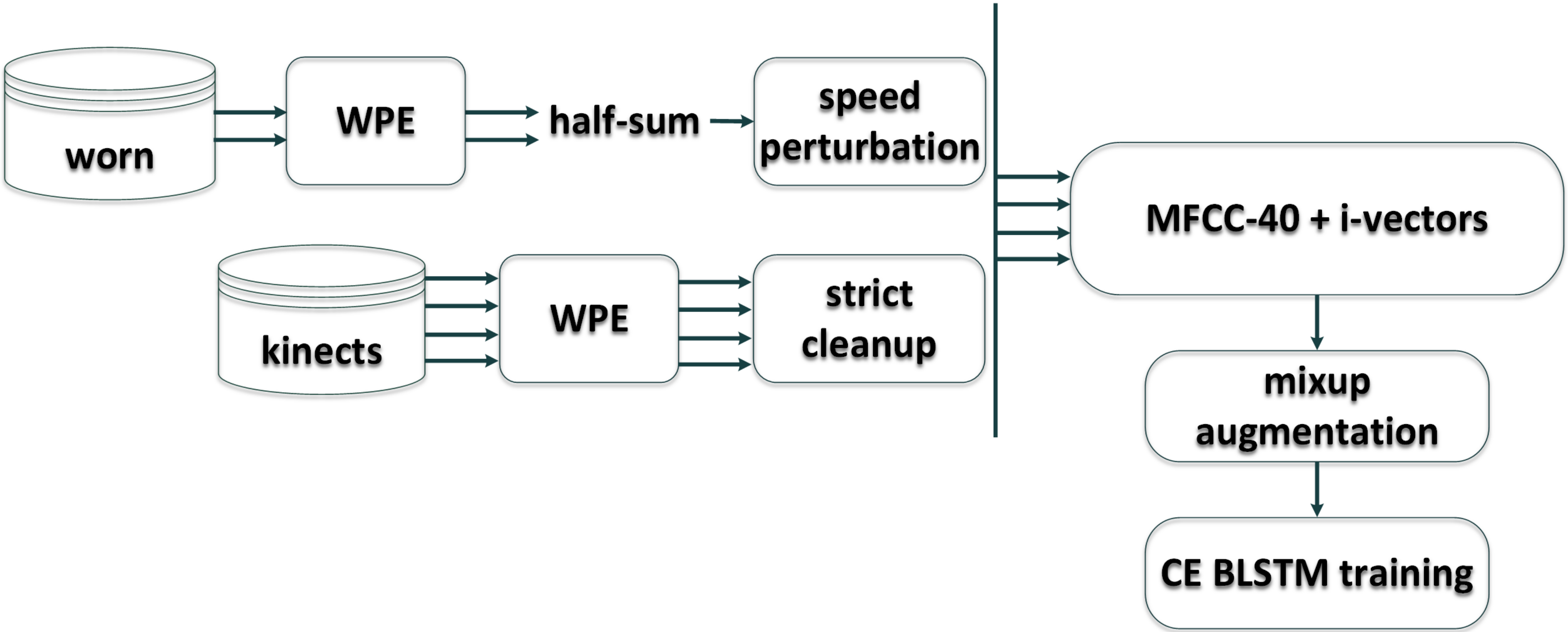


System II and III



Success story

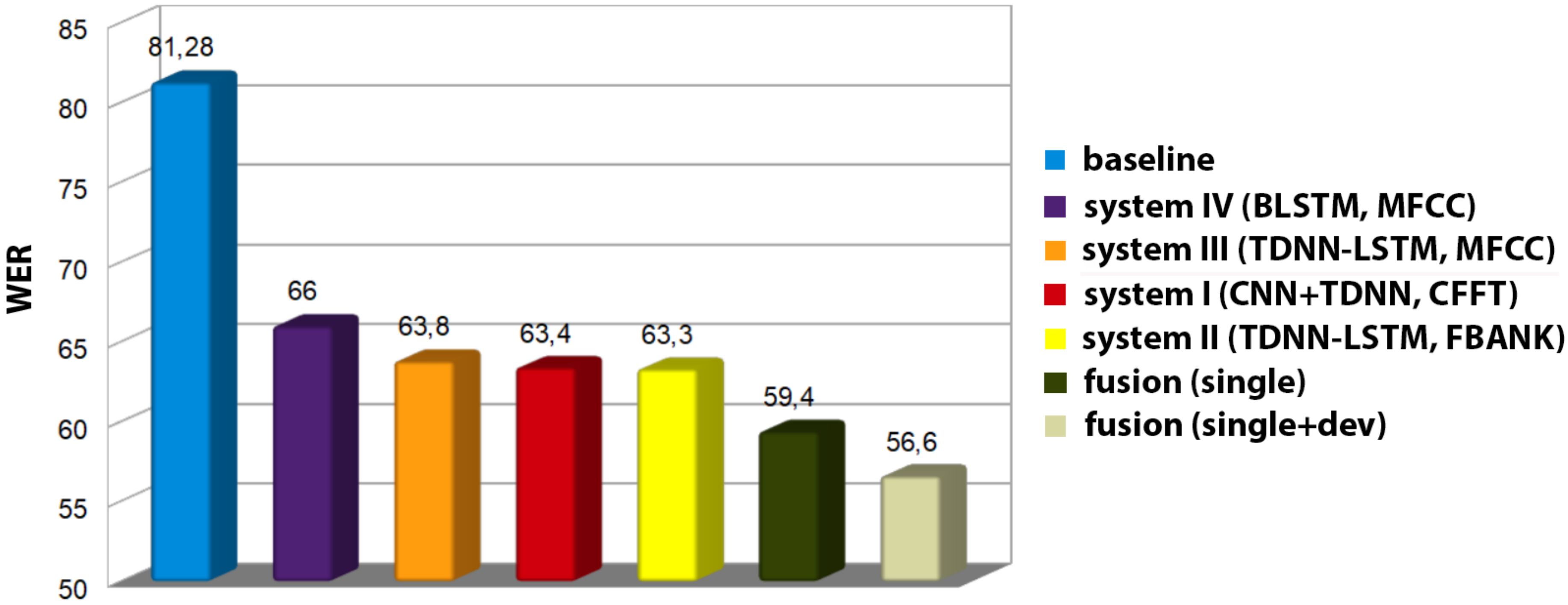
System IV



Decoding and models combination

- Decoding: application of softmax temperature to a prior distribution
- Fusion: posterior-level combination or two types of lattice-level combination

Fusion


85
80
75
70
65
60
55
50
WER
81,28
66
63,8
63,4
63,3
59,4
56,6
baseline
system IV (BLSTM, MFCC)
system III (TDNN-LSTM, MFCC)
system I (CNN+TDNN, CFFT)
system II (TDNN-LSTM, FBANK)
fusion (single)
fusion (single+dev)

WER (%) for the final system per session and location

Track	Session	Kitchen	Dining	Living	Overall
Single	S02 S09	67.7 58.0	59.7 59.8	55.5 54.9	59.4
Single+Dev	S02 S09	65.5 55.7	56.2 56.8	52.4 51.9	56.6
Multiple	S02 S09	65.8 55.5	57.9 57.3	55.1 55.4	58.1
Multiple+Dev	S02 S09	62.1 51.2	52.2 51.6	50.2 51.4	53.5

Summary

Track	Features	Adaptation	Model	Loss	WER
Single	CFFT FBANK MFCC MFCC	Auxiliary soft-mask soft-mask ivec+mask	CNN+TDNN TDNN-LSTM TDNN-LSTM BLSTM	CE, LF-MMI LF-MMI LF-MMI CE	63.4 63.3 63.8 66.0
	Fusion (4 systems)				59.4
Single+Dev	Fusion (4 systems)				56.6
Multiple	Fusion (4 systems)*				58.1
Multiple+Dev	Fusion (4 systems)*				53.5

Conclusions

Conclusions

- Common speech processing approaches face great challenges in real-world conditions

Conclusions

- Common speech processing approaches face great challenges in real-world conditions
- Both speaker separation and speaker adaptation are extremely important

Conclusions

- Common speech processing approaches face great challenges in real-world conditions
- Both speaker separation and speaker adaptation are extremely important
- Data augmentation and normalization are reasonably effective for this type of data

Conclusions

- Common speech processing approaches face great challenges in real-world conditions
- Both speaker separation and speaker adaptation are extremely important
- Data augmentation and normalization are reasonably effective for this type of data
- Fusion always gives a good performance improvement

STC innovations

Final results on eval and future work

Baseline	Our result	abs, %	rel, %
73.3	55.5	-17.8	-24.3

- Joint training of all components (front-end and back-end)
- Diarization for unsegmented real-world data

STC innovations

Contributions of applied methods

Method	Abs WER improvement, %
Array synchronization	0.9
Room simulator	1.6
Alignment transfer (worn half-sum → kinect)	1.3
Speaker adaptation (gating/throw out)	7/5
Speaker adaptation (i-vector)	2.4
Speaker adaptation (auxiliary)	4.1
Multi-channel model	2.2
Strict cleanup	1.3
WPE	1.4
Mixup	1.1
Speed Perturbation	0.9
Backstitch training	0.5
Fusion	3.9

THANK YOU

ABOUT THE COMPANY

STC-Innovations is a leader in the multimodal biometric market. STC-Innovations develops multimodal biometric solutions based on person-identifying technologies via voice, face and other noncontact biometric features.

STC-Innovations is a spin-off company of the Speech Technologies Center, leading global provider of innovative systems in high-quality recording, audio and video processing and analysis, speech synthesis and recognition, and real-time, high-accuracy voice and facial biometrics solutions with over 20 years of research, development and implementation experience in Russia and internationally.

STC is ISO-9001: 2008 certified.

CONTACTS

Russia
4 Krasutskogo street, St. Petersburg, 196084
Tel.: +7 812 325-8848
Fax: +7 812 327 9297
Email: info@speechpro.com

USA
Suite 316, 369 Lexington ave
New York, NY, 10017
Tel.: +1 646 237 7895
Email: sales-usa@speechpro.com