The Voice Conversion Challenge 2016

Tomoki Toda (Nagoya U, Japan) 

Ling-Hui Chen (USTC, China) 

Daisuke Saito (Tokyo U, Japan) 

Fernando Villavicencio (NII, Japan) 

Mirjam Wester (CSTR, UK) 

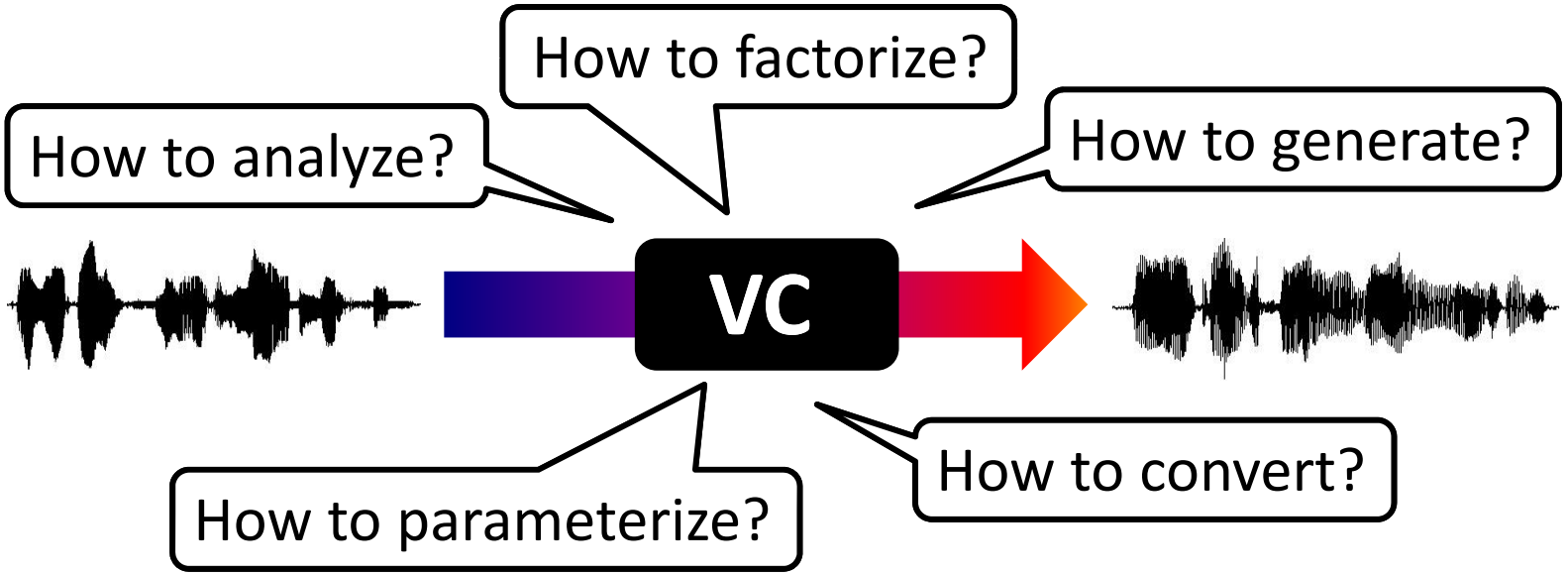
Zhizheng Wu (CSTR, UK) 

Junichi Yamagishi (NII/CSTR, Japan/UK) 

Sep. 10th, 2016

Voice Conversion (VC)

- Technique to modify speech waveform to **convert non-/para-linguistic information** while **preserving linguistic information**



- Research progress since **the late 1980s**
  - Development of various VC techniques (& potential applications)
  - **Not straightforward to compare across different VC techniques…**

Voice Conversion Challenge 2016

Objective

Better understand different VC techniques by comparing their performance using a freely-available dataset as a common dataset

- Following a policy of **Blizzard Challenge** **[Black & Tokuda, 2005]**

  "Evaluation campaign" rather than **"competition"**

- Also reveal a **risk** of VC techniques
  - Effective but possible to be used for spoofing
  - Important to inform people of VC as "kitchen knife"

Timelines of VCC 2016

(Sep. 9th, 2015) — (Short announcement at INTERSPEECH2015)

Nov. 18th, 2015 — Announcement & registration open

Nov. 25th, 2015 — **Release of training data**

1.5 months for training

Jan. 8th, 2016 — **Release of evaluation data**

1 week for conversion

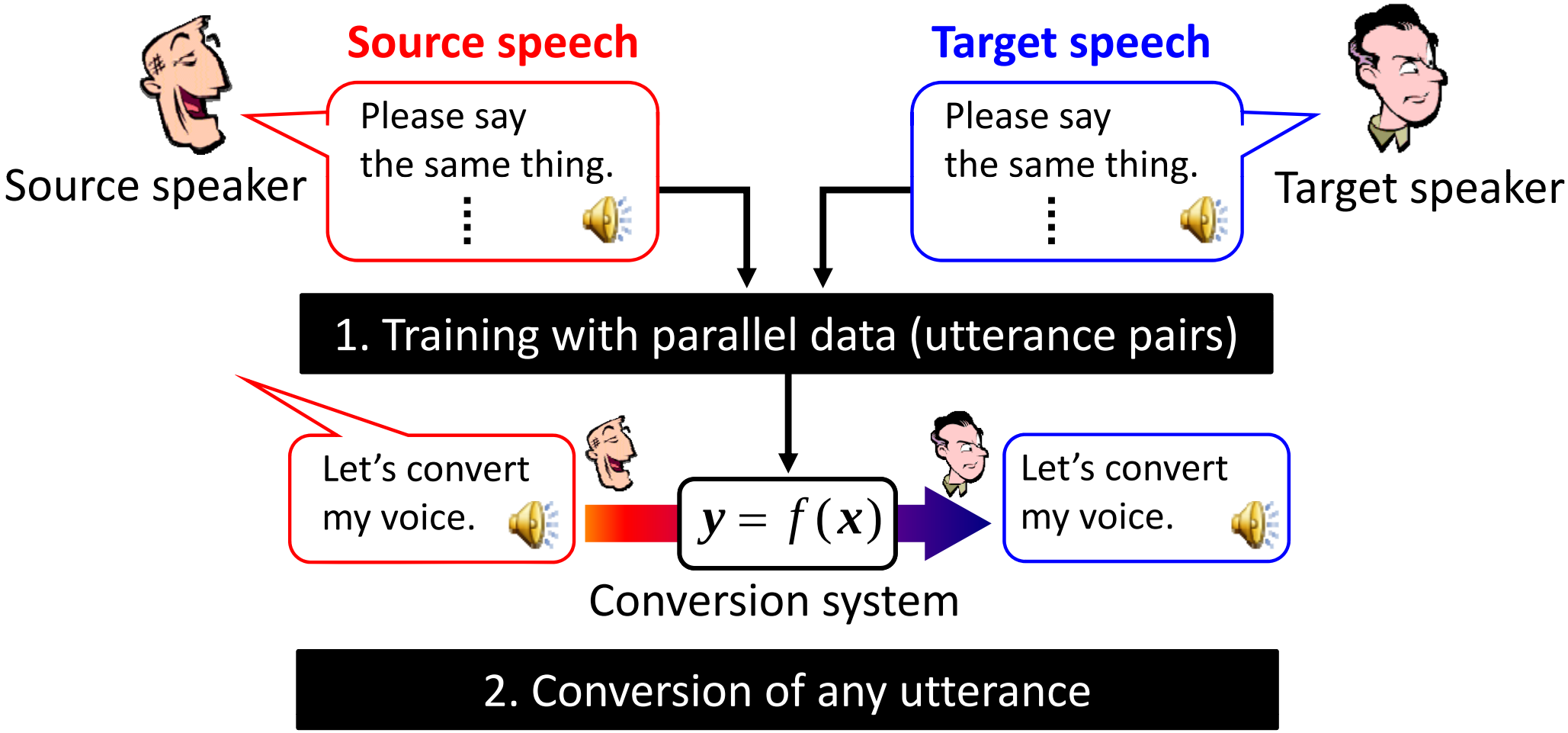
Jan. 15th, 2016 — **Deadline to submit the converted voice samples**

1.5 months for evaluation

Feb. 29th, 2016 — **Notification of results**

Task of VCC 2016

- Simple **speaker identity conversion** **[Abe *et al.*, 1990]**
  - Develop conversion systems using parallel data of each speaker pair



VCC 2016 Dataset [http://dx.doi.org/10.7488/ds/1430]

- **DAPS** (**D**ata **A**nd **P**roduction **S**peech) **[Mysore, 2015]**
  - Professional US English speakers
  - Freely available **[https://archive.org/details/daps_dataset]**
- Design of **VCC 2016 dataset**
  - Select 10 speakers including 5 female and 5 male speakers
  - Manually segmented into 216 sentences in each speaker
  - Down-sampled to 16 kHz

	# of speakers	**# of sentences**
Sources	3 females & 2 males	162 for training & 54 for evaluation
Targets	2 females & 3 males	162 for training

Rules of VCC 2016

- **Requirement**
  - Develop all 5 x 5 = 25 combinations of source-target pairs
- **Main guidelines**
  - Transform any acoustic features → **OK !**
  - Manual edit or tuning of systems in conversion → **NOT allowed**
  - Use manual transcriptions → **NOT allowed**
  - Use automatic speech recognition (ASR) → **OK!**
  - To develop a system for a certain speaker pair using data of other pairs within VCC 2016 dataset → **NOT allowed**
  - Use external data outside VCC 2016 dataset → **OK!**
  - Discard a part of utterances of the training set → **OK!**
  - Submit multiple entries → **NOT allowed**

Evaluation Methodology

- **Subjective evaluation**
  - Use only 16 speaker pairs (2 males & 2 females) from 25 speaker pairs
  - Use headphones in sound-treated booths
  - Listeners: 200 subjects

1. Opinion test on **naturalness**
   - Evaluate naturalness of each voice sample using a 5-scale opinion score
     - 1 (completely unnatural) to 5 (completely natural)

2. Pair-comparison test on **speaker similarity**
   - Judge whether 2 voice samples are uttered by the same speaker
     - Decision with confidence

Same, absolutely sure	**Same**, not sure	**Different**, not sure	**Different**, absolutely sure

Baseline System (Freely Available)

- **VC tools** **[Toda]** within **FestVox** **[Black & Lenzo]**
  - **Analysis methods**
    - F_0 extraction with **Edinburgh Speech Tools** **[Taylor *et al.*]**
    - Spectral analysis with **Signal Processing Toolkit (SPTK)** **[Tokuda *et al.*]**
  - **Converted parameters**
    - Mel-cepstrum (**MCEP**): Trajectory-wise conversion (**MLPG**) using global variance (**GV**) w/ Gaussian mixture model (**GMM**)
    - Log-scaled F_0 (**LF_0**): Linear transformation w/ mean & variance (**M&V**)
  - **Synthesis methods**
    - Simple pulse/noise excitation
    - Mel-log spectrum approximate (**MLSA**) filter

Submitted Systems

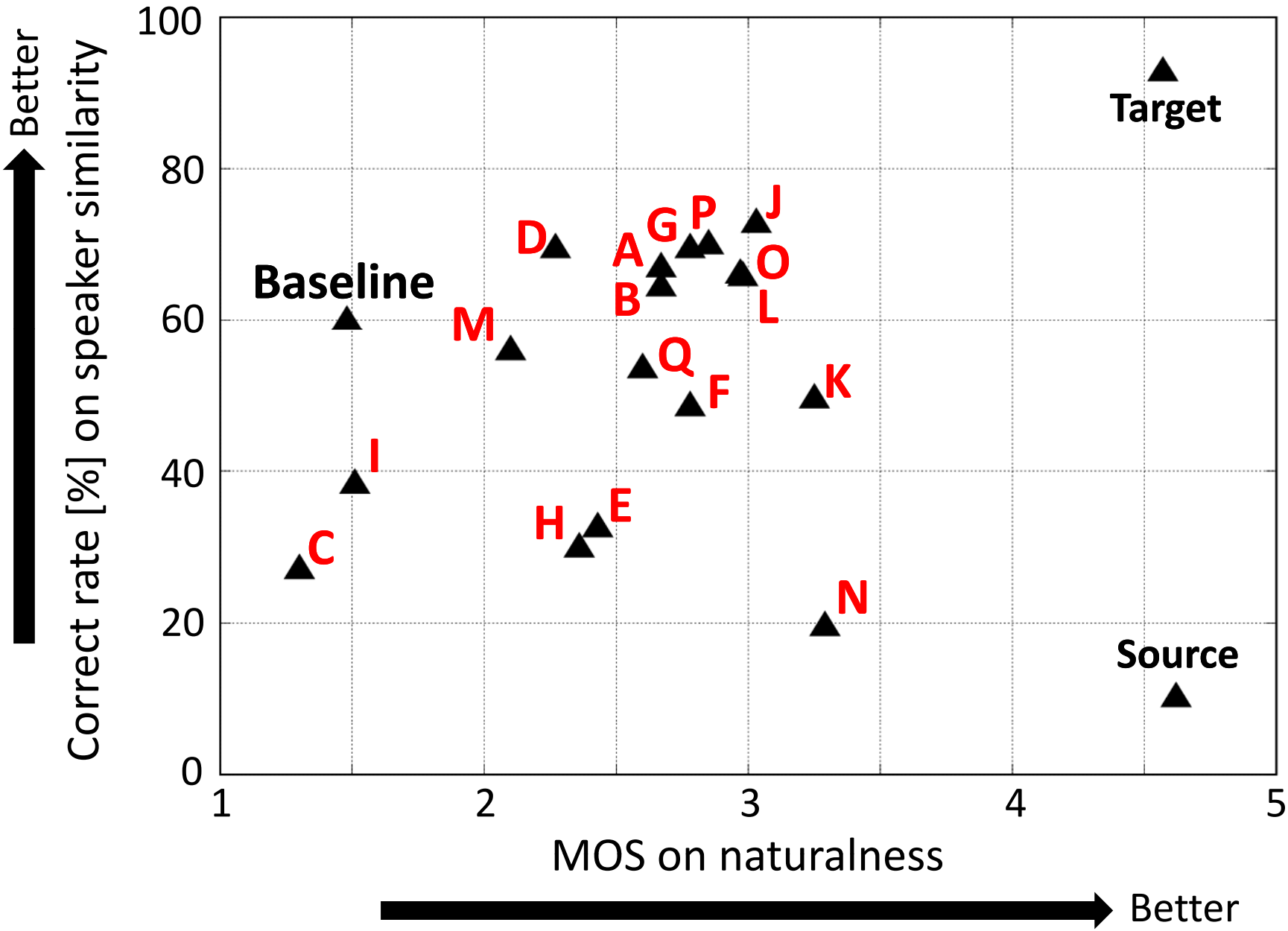
Team name	Ana-Syn	Converted Parameters & Conversion Methods					ASR	+DB
A	Ahocoder	MCEP	GMM, **MGE**, MLPG, PF	LF_0 M&V			No	No
B	STRAIGHT	MCEP	Exemplar, MLPG, GV	LF_0 M&V			No	No
C	STRAIGHT	MLSP	DNN & GMM, PF	LF_0 M&V			No	Yes
D	STRAIGHT	MCEP	MDN & GMM, PF	LF_0 M&V			No	No
E	Ahocoder	MCEP	GMM, **FW** & **Scaling**	LF_0 M&V			No	No
F	STRAIGHT	MCEP	Phone posteriorgram	LF_0 M&V			Yes	Yes
G	STRAIGHT	MCEP	LSTM-RNN	LF_0 M&V		Spk rate	Yes	Yes
H	STRAIGHT	MCEP	DNN, MTL	LF_0 M&V		Spk rate	Yes	Yes
I	Ahocoder	LSP	GMM, MMSE, **i-vector**	LF_0 M&V			No	Yes
J	STRAIGHT	MCEP	GMM, MS, **diff filter**	LF_0 M&V	BAP		No	No
K	TEAP	MLSP	**FW** & GMM, **diff filter**	F_0 shift		Spk rate	No	No
L	STRAIGHT	**Multi systems** & **selection**		LF_0 M&V	Resid		Yes	Yes
M	STRAIGHT	MCEP	LSTM	LF_0 M&V			No	No
N	LPC	LP coef	FW	F_0 shift		Spk rate	No	No
O	STRAIGHT	ST spec	**FW** & GTDNN	LF_0 LSTM	BAP		No	No
P	STRAIGHT	MCEP	GMM, MLPG, GV	LF_0 M&V	BAP		No	No
Q	Ahocoder	MCEP	Frame selection, MLPG	LF_0 M&V			No	No

Submitted Systems

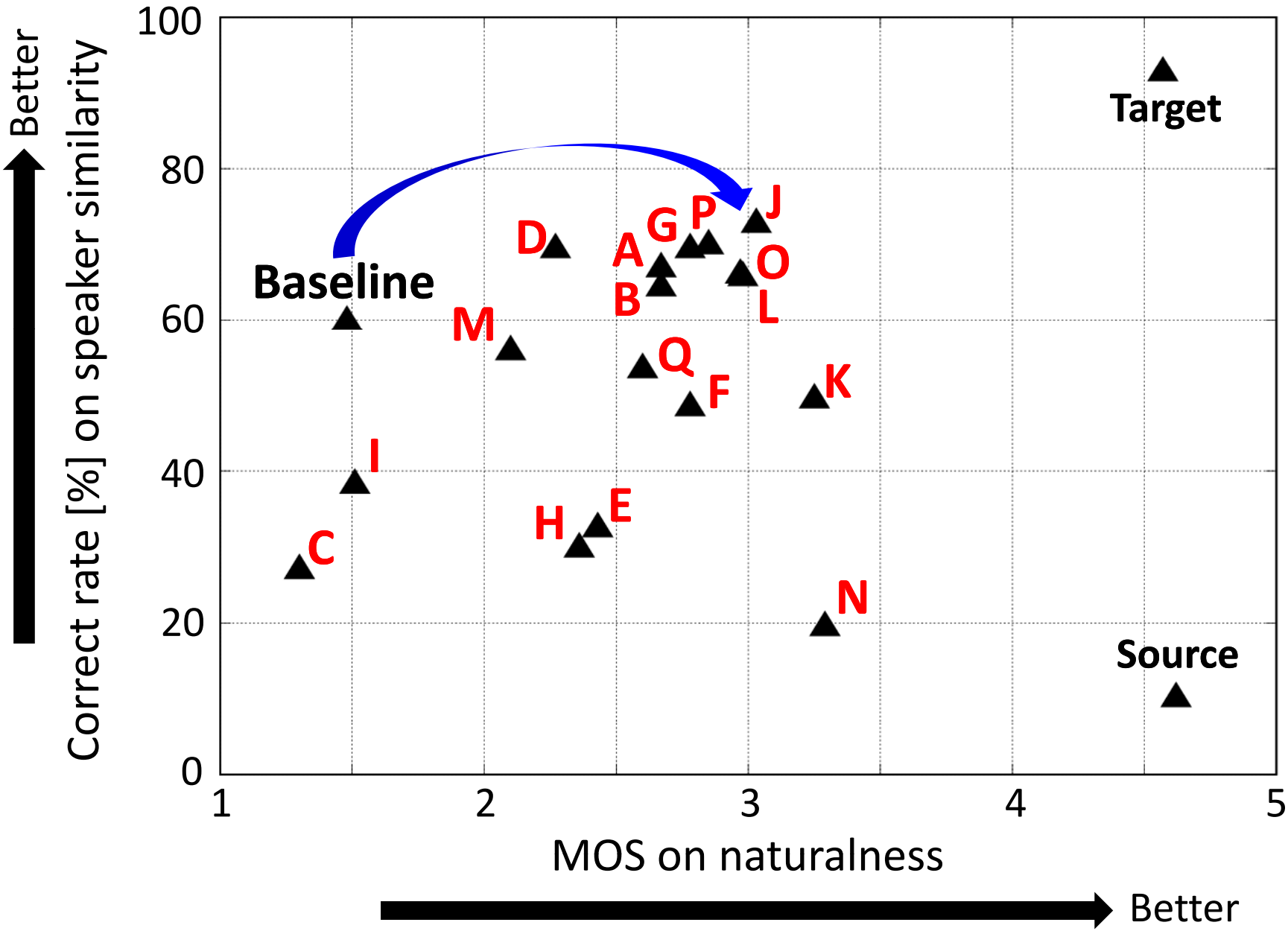
Spectral envelope | F_0 pattern | Excitation | Duration

Team name	Ana-Syn	Converted Parameters & Conversion Methods					ASR	+DB
A	Ahocoder	MCEP	**GMM**, **MGE**, MLPG, PF	LF_0 M&V			No	No
B	STRAIGHT	MCEP	**Exemplar**, MLPG, GV	LF_0 M&V			No	No
C	STRAIGHT	MLSP	**DNN** & **GMM**, PF	LF_0 M&V			No	Yes
D	STRAIGHT	MCEP	**MDN** & **GMM**, PF	LF_0 M&V			No	No
E	Ahocoder	MCEP	**GMM**, **FW** & **Scaling**	LF_0 M&V			No	No
F	STRAIGHT	MCEP	**Phone posteriorgram**	LF_0 M&V			Yes	Yes
G	STRAIGHT	MCEP	**LSTM-RNN**	LF_0 M&V		Spk rate	Yes	Yes
H	STRAIGHT	MCEP	**DNN**, MTL	LF_0 M&V		Spk rate	Yes	Yes
I	Ahocoder	LSP	**GMM**, MMSE, **i-vector**	LF_0 M&V			No	Yes
J	STRAIGHT	MCEP	**GMM**, MS, **diff filter**	LF_0 M&V	BAP		No	No
K	TEAP	MLSP	**FW** & **GMM**, **diff filter**	F_0 shift		Spk rate	No	No
L	STRAIGHT	**Multi systems** & **selection**		LF_0 M&V	Resid		Yes	Yes
M	STRAIGHT	MCEP	**LSTM**	LF_0 M&V			No	No
N	LPC	LP coef	FW	F_0 shift		Spk rate	No	No
O	STRAIGHT	ST spec	**FW** & **GTDNN**	LF_0 LSTM	BAP		No	No
P	STRAIGHT	MCEP	**GMM**, MLPG, GV	LF_0 M&V	BAP		No	No
Q	Ahocoder	MCEP	**Frame selection**, MLPG	LF_0 M&V			No	No

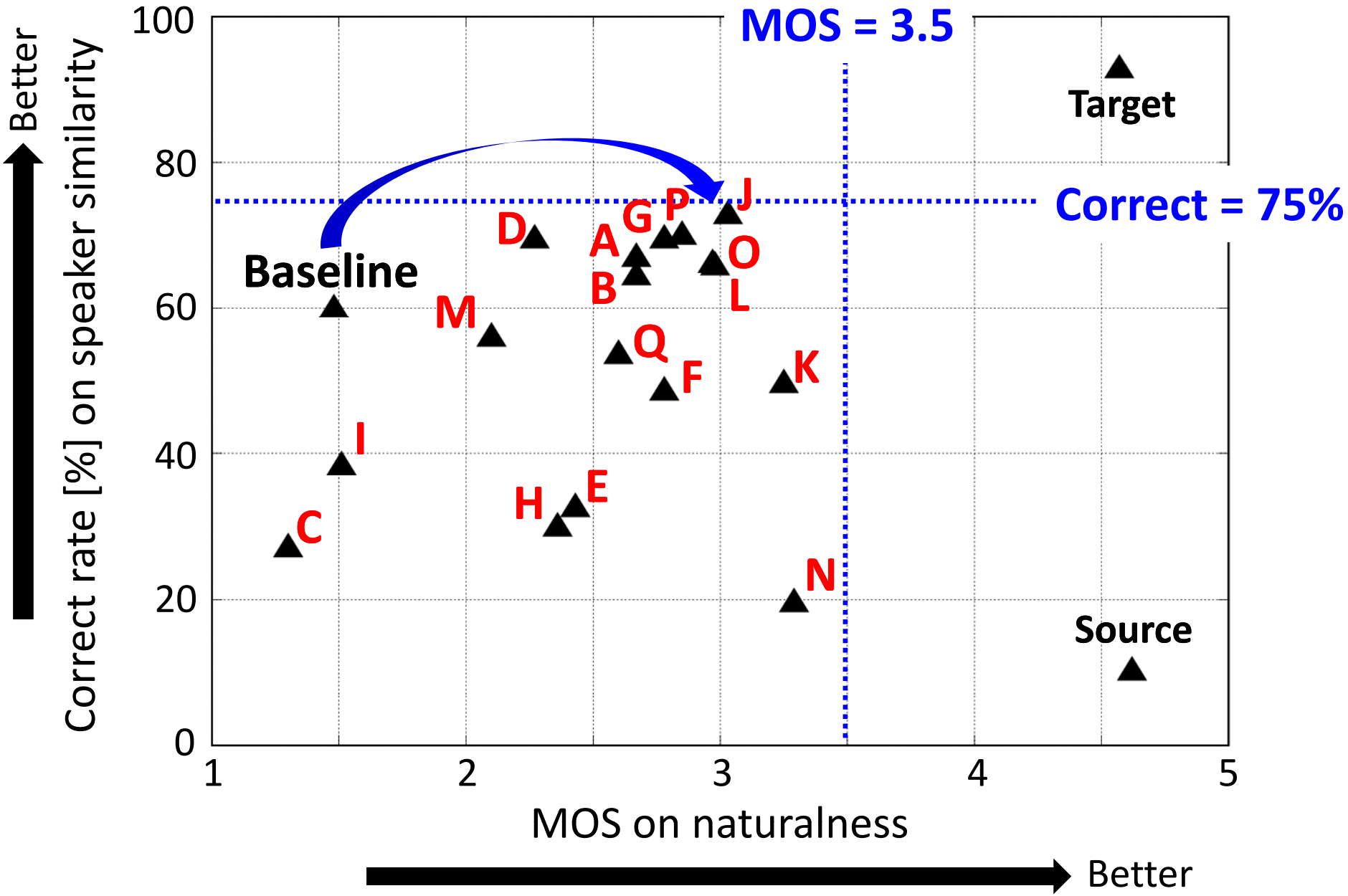
Overall Results of Listening Tests



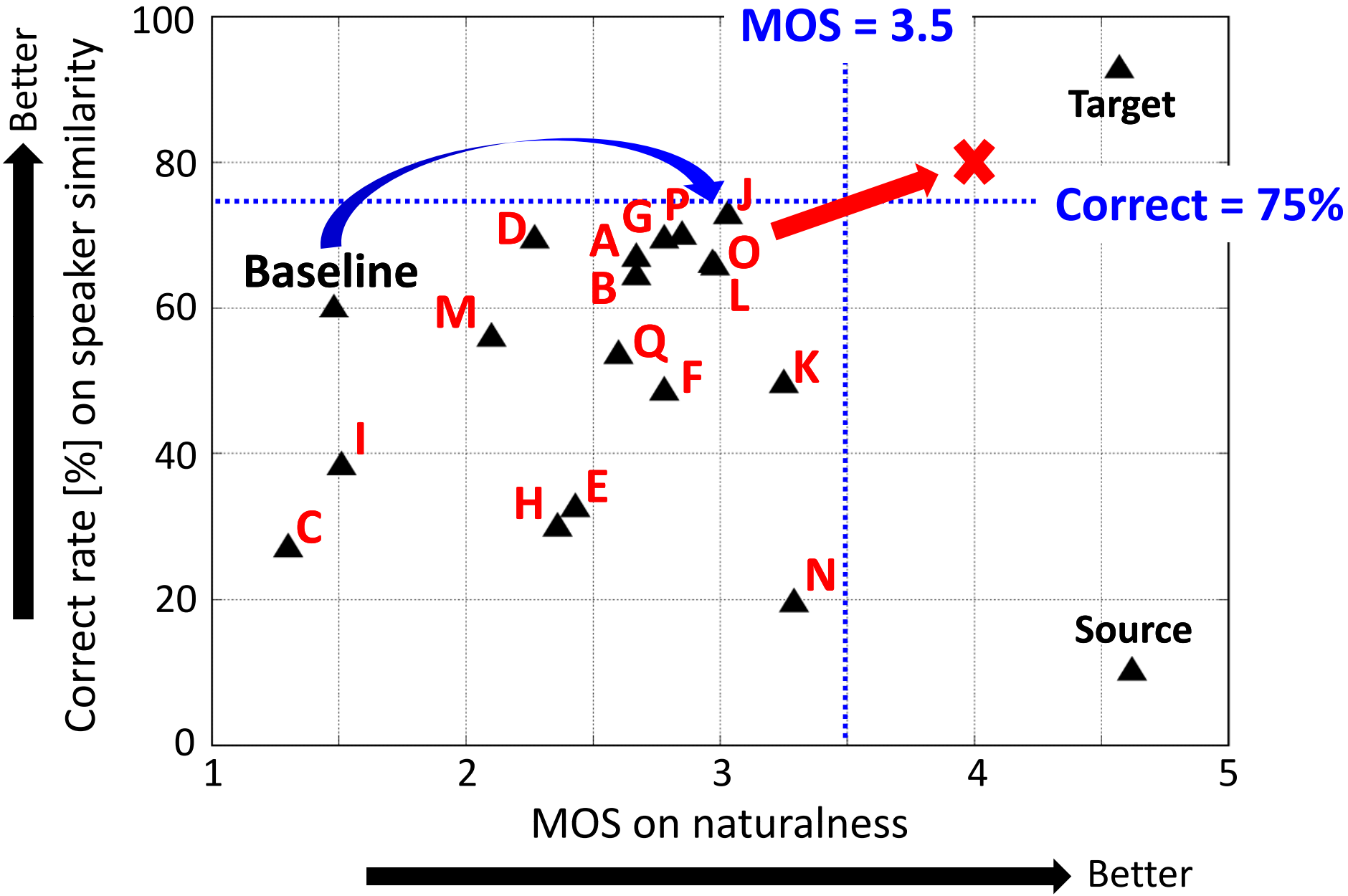
Overall Results of Listening Tests



Overall Results of Listening Tests



Overall Results of Listening Tests



Discussion and Future Plan

- **Issues of listening test**
  - US English evaluated by British English subjects (less sensitive to prosody?)
  - Hard to separately evaluate prosodic and spectral conversion
- **Suggestions towards next challenge**
  - Use fewer or more training utterances
  - Use non-parallel datasets
  - Use data recorded in non-ideal acoustic conditions
- **Future plan and collaboration**
  - Provide converted voices for the Automatic Speaker Verification Spoofing and Countermeasures (**ASVspoof**) Challenge **[Wu *et al.*, 2015]**
  - **Hold VCC every 2 years (?)**
  - **Appreciate you help** (*e.g.*, provide data, manage evaluation, …)***!***

Conclusions

- **Voice Conversion Challenge 2016 (VCC 2016)**
  - Task: **speaker identity conversion**
  - Datasets: **VCC 2016 dataset** from DAPS dataset
  - Participants: **17 teams**
  - Test: **naturalness** & **speaker similarity** evaluated by **200 subjects**
  - Results: **MOS on naturalness < 3.5** & **correct rate on similarity < 75%**

✓ **VCC homepage**: **http://vc-challenge.org/** (to be updated)

✓ **Datasets & results**: **http://dx.doi.org/10.7488/ds/1430**

✓ **Email: vcc2016@vc-challenge.org**

Any comments and suggestions are very welcome!

Acknowledgement

- We are grateful to
  - **COLIPS** for sponsoring the evaluation
  - **iFLYTEK** for supporting database development
- This work was supported in part by
  - **EPSRC through Programme Grant EP/I031022/1 (NST) and EP/J002526/1 (CAF)**
  - **JSPS KAKENHI Grant Number 26280060**
- We specially thank to **Blizzard Challenge Organizers [King *et al.*]** to kindly allow us to use the evaluation system!