SP-L8.2

Innovative R&D by NTT

Soft-Target Training with Ambiguous Emotional Utterances for DNN-based Speech Emotion Classification

NTT Media Intelligence Laboratories, NTT Corporation

Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, Yushi Aono

ICASSP 2018

Summary

Purpose

✓ Speech emotion classification from acoustic features

– Task: 4-class classification (*Neutral, Happy, Sad, Angry*)

Novelty

✓ To mitigate training data limitation problem, utilizing ***ambiguous emotional utterances*** **(no target emotions are dominant) which are ignored in the conventional methods**

– Employ two types of soft-target training

Results

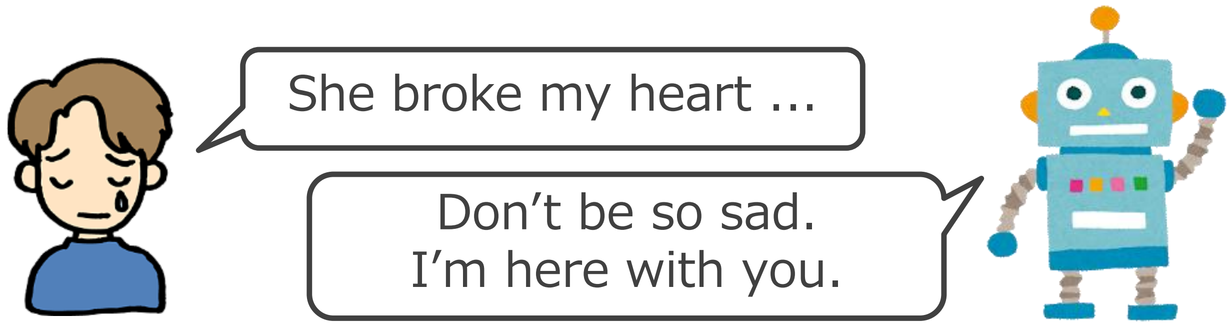
✓ Performance improved

– Overall Accuracy: 58.6% → 62.6%, Average Recall: 53.7% → 63.7%

Background

Speech emotion recognition is important technology to understand natural speech

- ✓ Application : **"sympathetic"** spoken dialog system

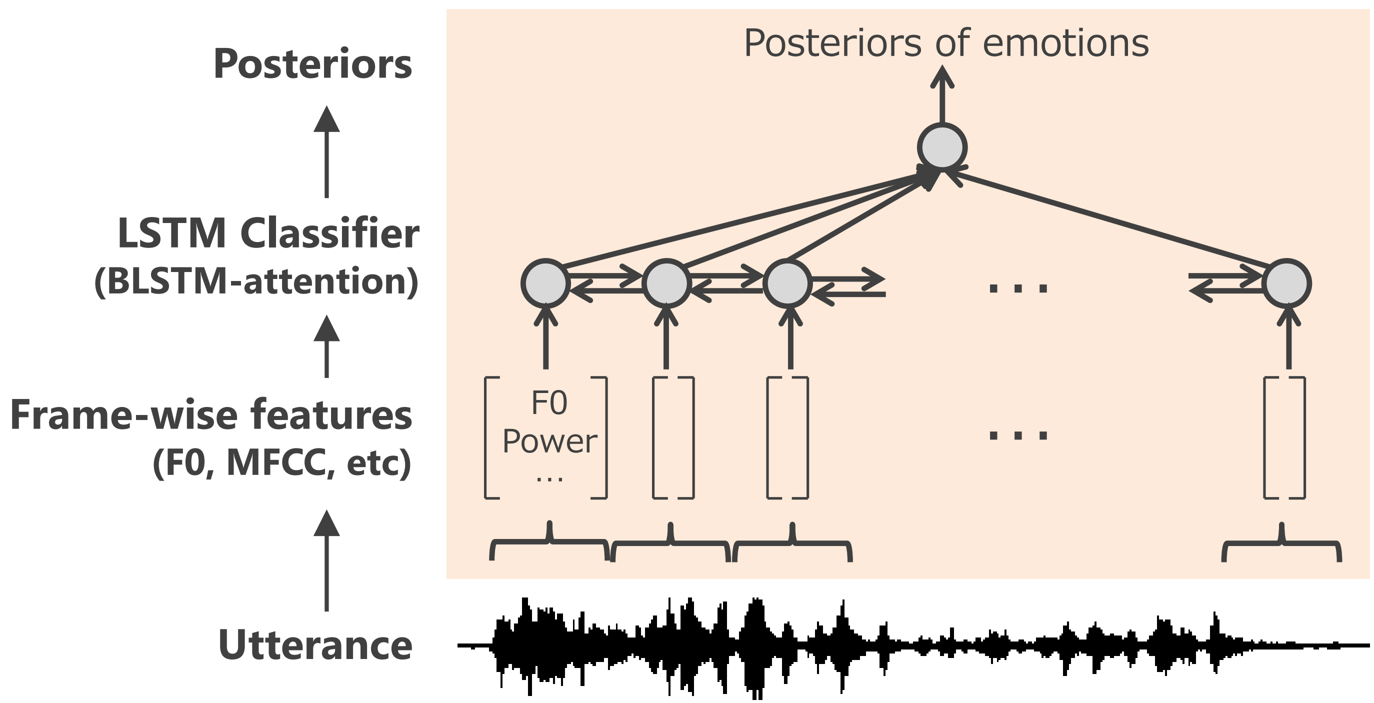


- ✓ Task description
  - – Input : short utterance (1～10 sec.)
  - – Target : 4-class speech emotion (*Neutral, Happy, Sad, Angry*)

Conventional

Frame-wise acoustic features + BLSTM-RNNs

- ✓ Emotion classification by BLSTM w/ attention [Mirsamadi+, 17]
  - – Utilizing **local characteristics** of emotions

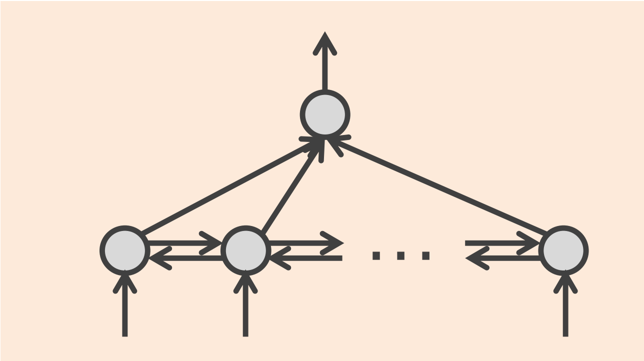


Problem

Training data is usually limited

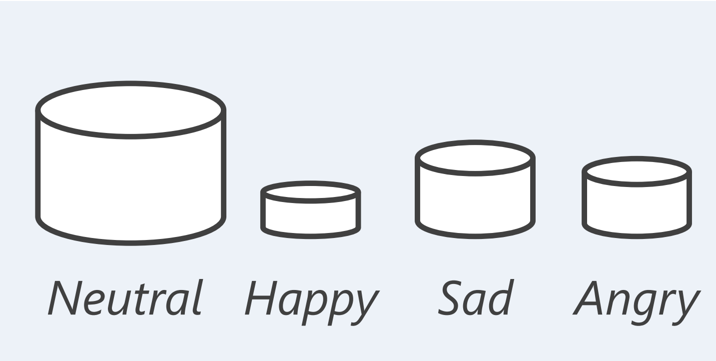
✓ Emotion classification by BLSTM w/ attention [Mirsamadi+, 17]

\# of parameters: **100k~**



↔

\# of train data: **~5k**



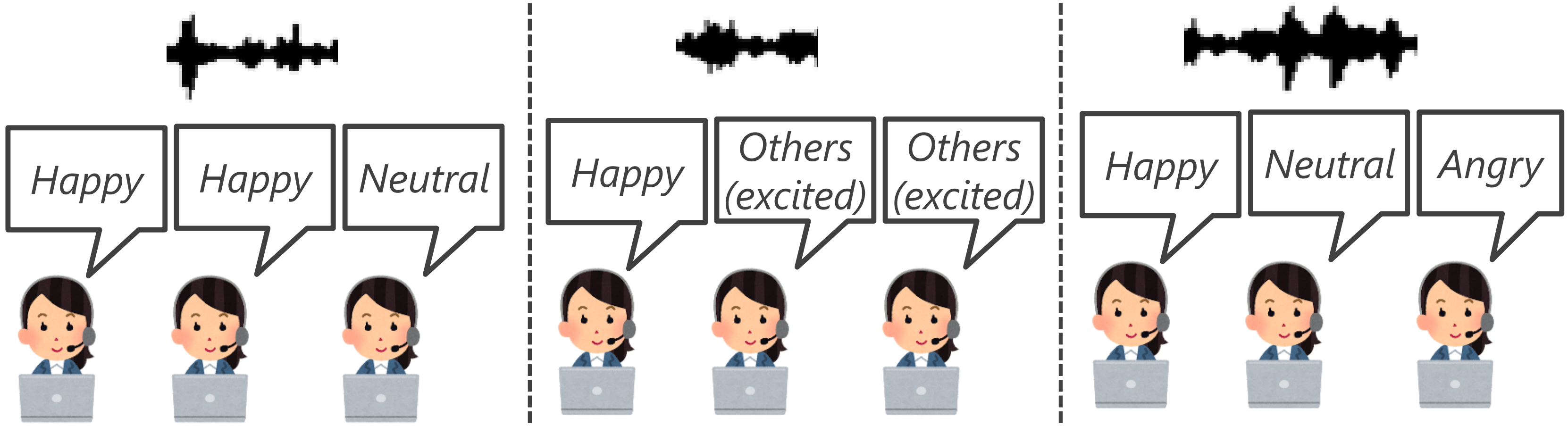
→ Classifier is overfitted / less generalized

<u>Issue</u> How to train complex classifier from limited data ?

Problem - Why limited?

Ground truths are decided by several annotators.
Some utterances are ignored for training

✓ Ground truth = **Dominant emotion** of annotations

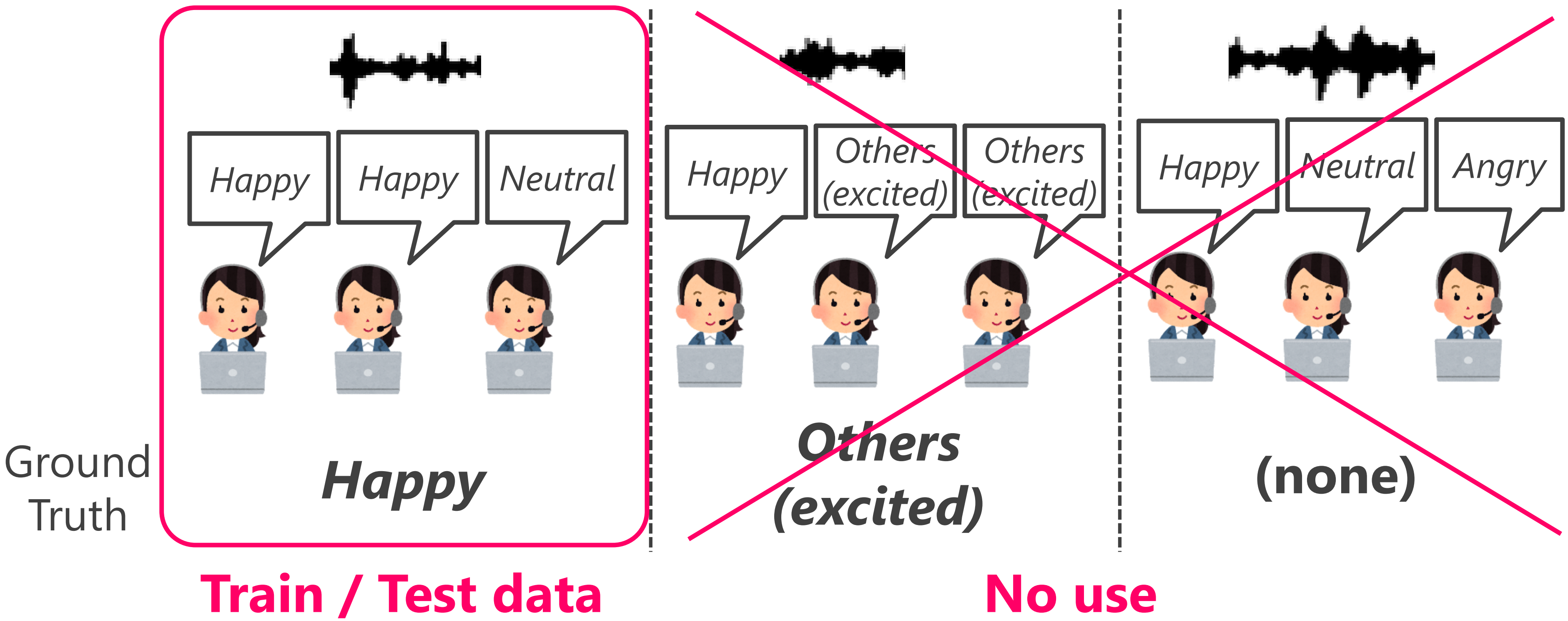


Ground Truth	***Happy***	***Others (excited)***	**(none)**

Problem - Why limited?

Ground truths are decided by several annotators.
Some utterances are ignored for training

- ✓ Ground truth = **Dominant emotion** of annotators



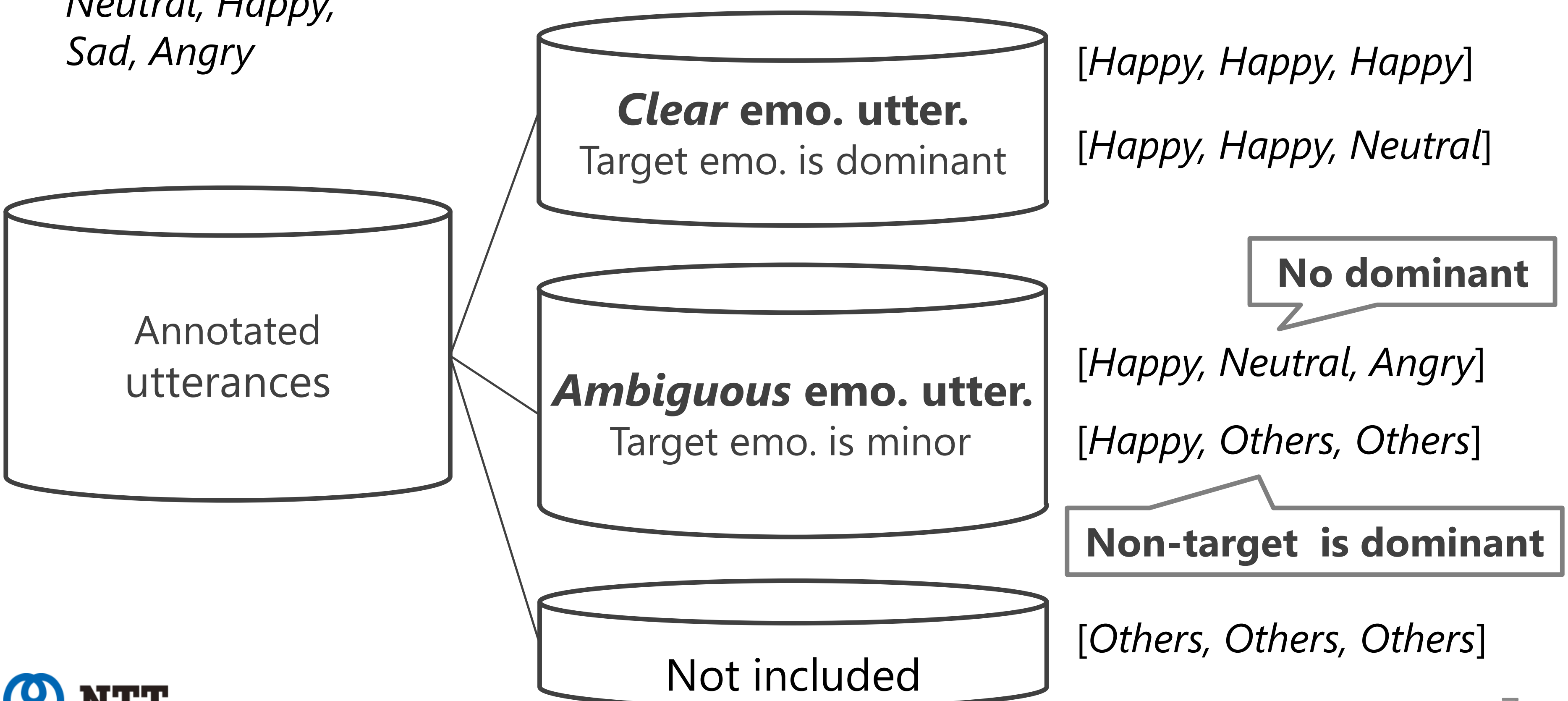
Approach (1/2)

Utilize *ambiguous* emotional utterances (target emo. are minor) to mitigate training data limitation

Target emotions

Neutral, Happy,
Sad, Angry

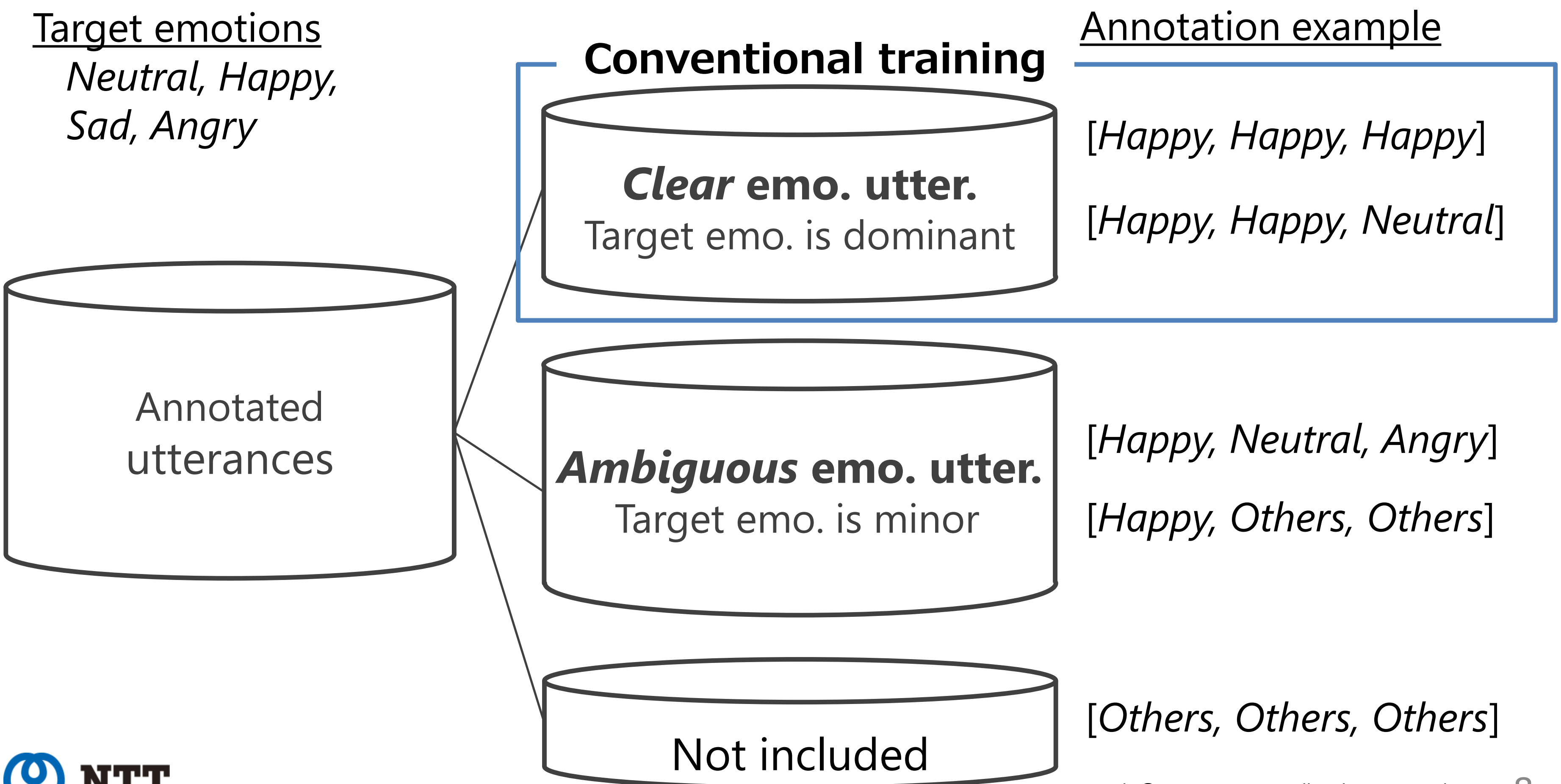
Annotation example



Approach (1/2)

Utilize *ambiguous* emotional utterances (target emo. are minor)
to mitigate training data limitation

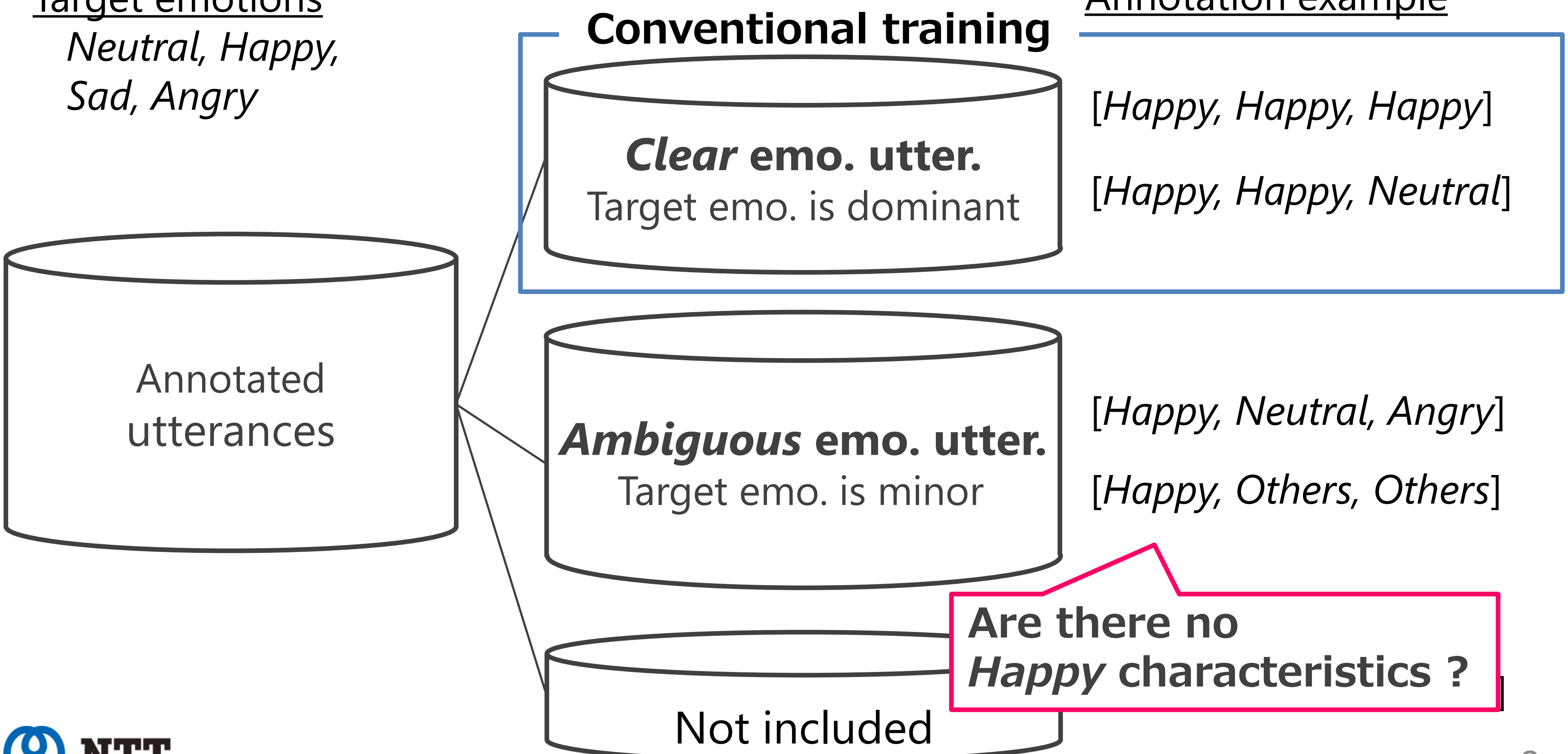
Target emotions
Neutral, Happy,
Sad, Angry



Approach (1/2)

Utilize *ambiguous* emotional utterances (target emo. are minor)
to mitigate training data limitation

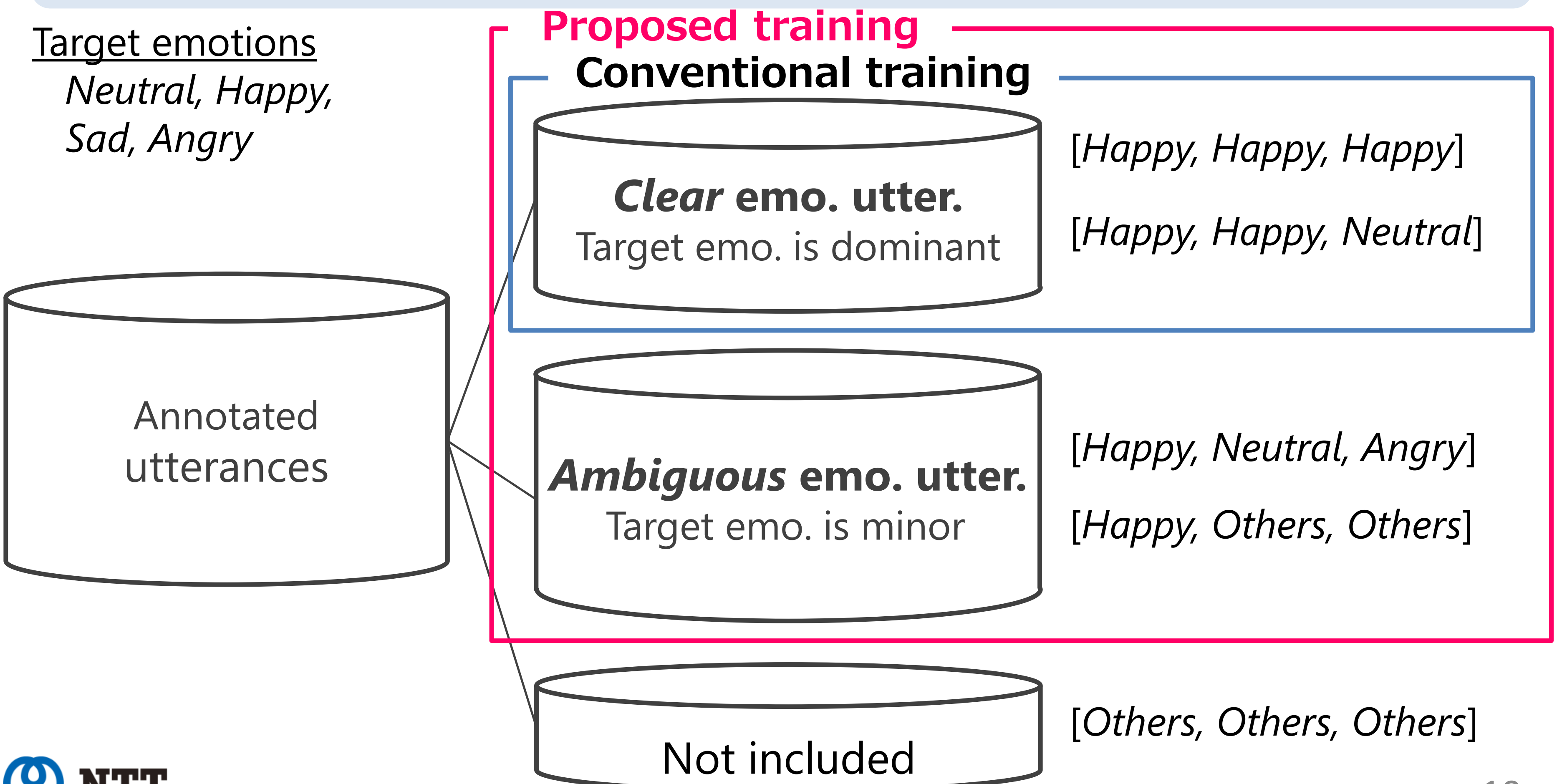
<u>Target emotions</u>
Neutral, Happy,
Sad, Angry



Approach (1/2)

Utilize *ambiguous* emotional utterances (target emo. are minor)
to mitigate training data limitation

Target emotions
Neutral, Happy,
Sad, Angry



Approach (2/2)

Control discriminativity to handle both *clear* and *ambiguous* emotional utterances effectively

High discriminativity
Train as
definitely *Happy*

***Clear* emo. utter.**
Target emo. is dominant

[*Happy, Happy, Happy*]

[*Happy, Happy, Neutral*]

Low discriminativity
Train as
maybe *Happy*

***Ambiguous* emo. utter.**
Target emo. is minor

[*Happy, Neutral, Angry*]

[*Happy, Others, Others*]

Not included

[*Others, Others, Others*]

Proposed

Soft-target training is employed to deal *clear/ambiguous* emotional utterances

✓ Two types of soft-target

1. Soft-target [Fayek+, 16]

$$q(c_k) = \frac{\sum_n h_k^{(n)}}{\sum_k \sum_n h_k^{(n)}}$$

Annotation frequency (sum=1)

$h_k^{(n)}$: Binary label-existence (0/1)
n-th annotator, *k*-th emotion class

K : Total emotion classes

2. Modified soft-target

$$q(c_k) = \frac{\alpha + \sum_n h_k^{(n)}}{\alpha K + \sum_k \sum_n h_k^{(n)}}$$

Additive smoothed form of conventional soft-target

α : Smoothing coefficient

✓ Model parameters are updated by cross-entropy loss

$$L = -\sum_{k=1}^{K} q(c_k) \log p(c_k | \mathbf{X}, \theta)$$

Proposed: modified soft-target

Modified soft-target is suitable to represent *ambiguous* emotional utterances

✓ Examples of teachers $q(c_k)$

	Hard-target	**Soft-target** [Fayek+,16]	**Modified Soft-target**
[Happy, Happy, Happy]	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.14, *Hap* 0.58, *Sad* 0.14, *Ang* 0.14
[Happy, Happy, Neutral]	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.33, *Hap* 0.66, *Sad* 0, *Ang* 0	*Neu* 0.29, *Hap* 0.43, *Sad* 0.14, *Ang* 0.14
[Happy, Others, Others]	(no use)	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.2, *Hap* 0.4, *Sad* 0.2, *Ang* 0.2

Non-target

(Smoothing coeff. $\alpha = 1$)

Proposed: modified soft-target

Modified soft-target is suitable to represent *ambiguous* emotional utterances

✓ Examples of teachers $q(c_k)$

	Hard-target	**Soft-target** [Fayek+,16]	**Modified Soft-target**
[*Happy, Happy, Happy*]	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.14, *Hap* 0.58, *Sad* 0.14, *Ang* 0.14
[*Happy, Happy, Neutral*]	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.33, *Hap* 0.66, *Sad* 0, *Ang* 0	*Neu* 0.29, *Hap* 0.43, *Sad* 0.14, *Ang* 0.14
[*Happy, Others, Others*]	(no use)	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.2, *Hap* 0.4, *Sad* 0.2, *Ang* 0.2

Non-target

Ambiguous utterances are discarded

(Smoothing coeff. $\alpha = 1$)

Proposed: modified soft-target

Modified soft-target is suitable to represent *ambiguous* emotional utterances

✓ Examples of teachers $q(c_k)$

	Hard-target	**Soft-target** [Fayek+,16]	**Modified Soft-target**
[*Happy, Happy, Happy*]	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.14, *Hap* 0.58, *Sad* 0.14, *Ang* 0.14
[*Happy, Happy, Neutral*]	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.33, *Hap* 0.66, *Sad* 0, *Ang* 0	*Neu* 0.29, *Hap* 0.43, *Sad* 0.14, *Ang* 0.14
[*Happy, Others, Others*]	(no use)	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.2, *Hap* 0.4, *Sad* 0.2, *Ang* 0.2

Non-target

Allocate same teacher labels to *clear/ambiguous*

(Smoothing coeff. $\alpha = 1$)

Proposed: modified soft-target

Modified soft-target is suitable to represent *ambiguous* emotional utterances

✓ Examples of teachers $q(c_k)$

	Hard-target	**Soft-target** [Fayek+,16]	**Modified Soft-target**
[*Happy, Happy, Happy*]	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.14, *Hap* 0.58, *Sad* 0.14, *Ang* 0.14
[*Happy, Happy, Neutral*]	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.33, *Hap* 0.66, *Sad* 0, *Ang* 0	*Neu* 0.29, *Hap* 0.43, *Sad* 0.14, *Ang* 0.14
[*Happy, Others, Others*]	(no use)	*Neu* 0, *Hap* 1.0, *Sad* 0, *Ang* 0	*Neu* 0.2, *Hap* 0.4, *Sad* 0.2, *Ang* 0.2

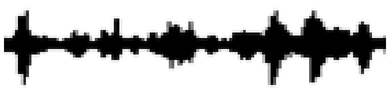
Non-target

Lower discriminativity in ambiguous emo. uttr.

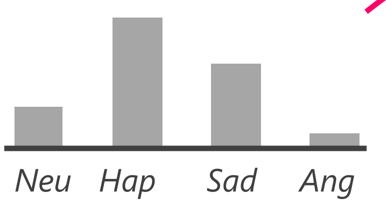
Interpretation

Modified soft-target is regarded as
Maximum a posteriori (MAP) estimation from annotations

Utterance



"true" distribution of target emo.



Annotations

Sampling
(N=# of annotations)

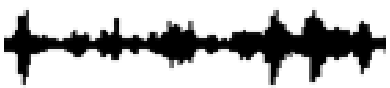
[*Happy, Happy, Sad*]

Objective function of the model

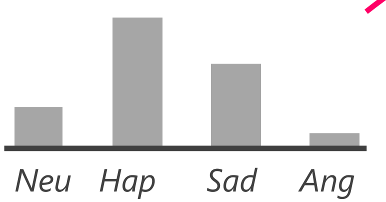
Interpretation

Modified soft-target is regarded as
Maximum a posteriori (MAP) estimation from annotations

Utterance



"true" distribution
of target emo.



Annotations

Sampling
(N=# of annotations)

[*Happy, Happy, Sad*]

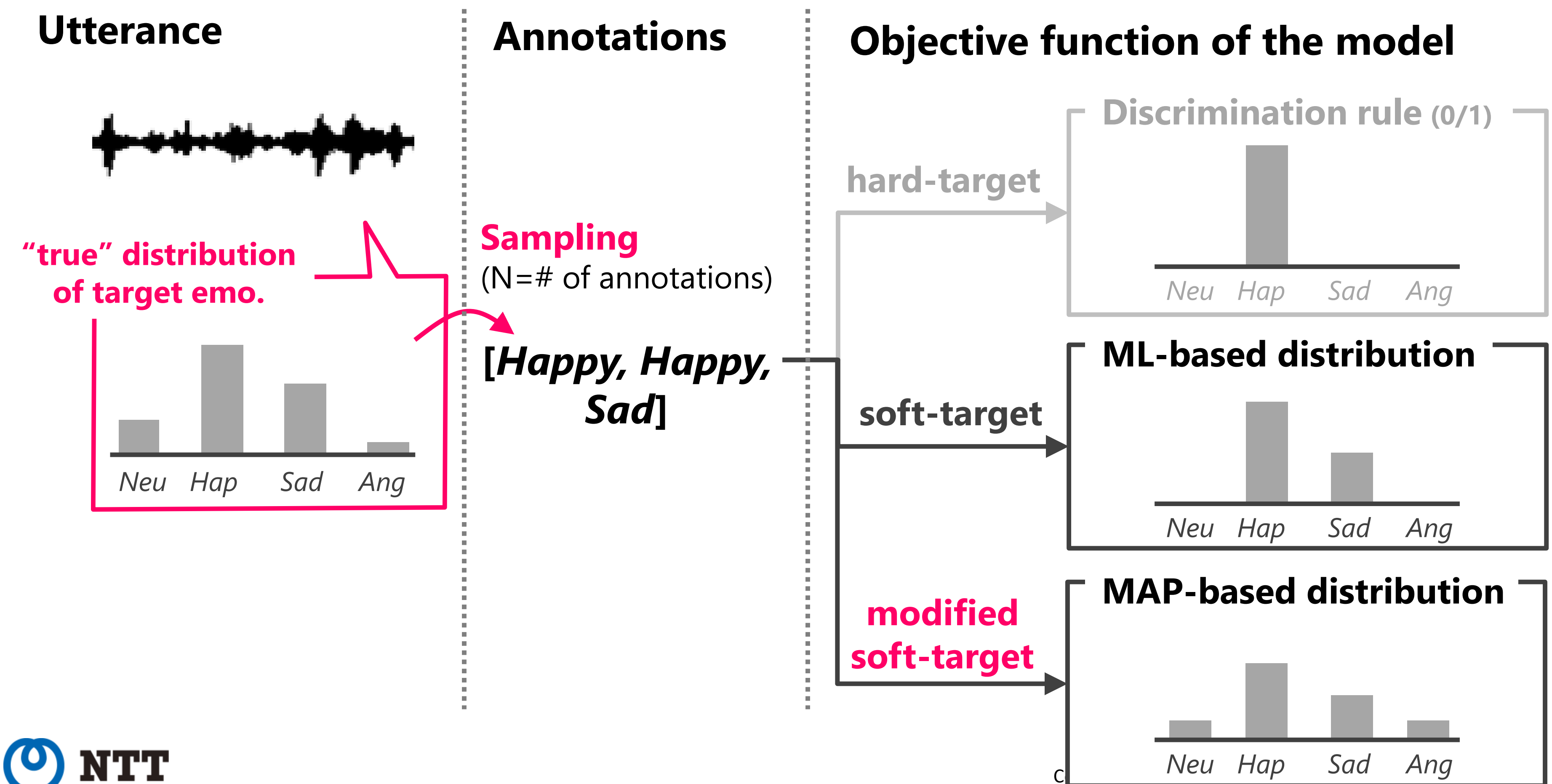
Objective function of the model

hard-target

Discrimination rule (0/1)

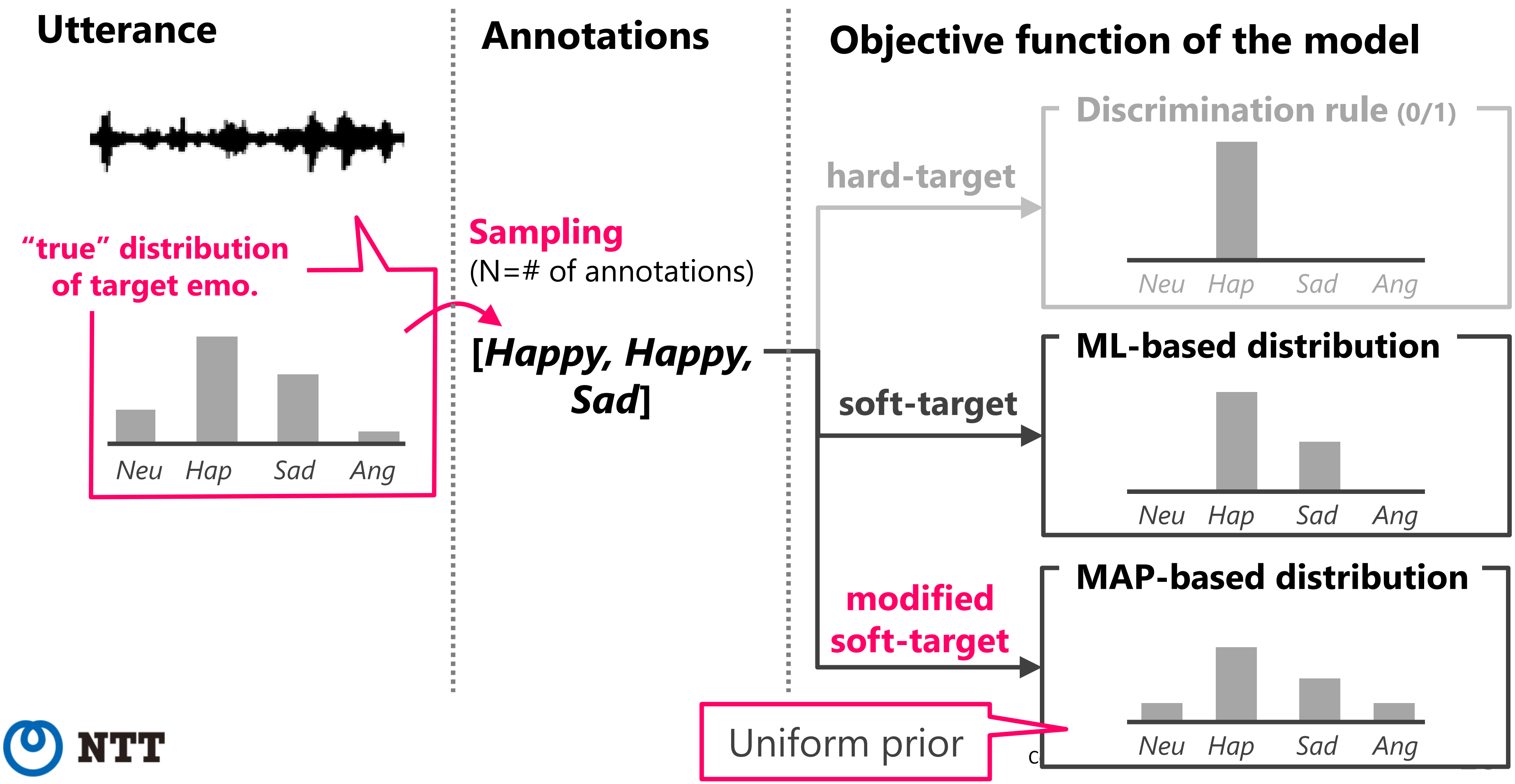
Interpretation

Modified soft-target is regarded as
Maximum a posteriori (MAP) estimation from annotations



Interpretation

Modified soft-target is regarded as
Maximum a posteriori (MAP) estimation from annotations



Experiments

- ✓ **Purpose**
  1. Evaluate effectiveness of *ambiguous* emotional utterances for train
  2. Compare teacher labels (hard / soft / modified soft)

- ✓ **Dataset:** IEMOCAP [Busso+, 08]
  - **Task:** 2-speaker dialogue (1 male, 1 female)
  - **# of speakers:** 10 (train: 8, test: 2)
  - **# of annotators:** 3

frustrated, *excited*, *surprised*, *fear*, *disgust*, no-dominant

		Total	# of utterances (dominant emotion)				
			Neutral	*Happy*	*Sad*	*Angry*	*Others*
Train	*clear*	3548	1324	460	890	874	-
	ambiguous	3693	0	0	0	0	3693
Test		942	384	135	194	229	-

Setups

- ✓ **Classifier:** BLSTM + attention [Mirsamadi+,17]
  - **Structure**
    - ➢ Full256-BLSTM128-attention-Full256
  - **Input:** frame-wise acoustic features, 47 dims.
    - ➢ MFCC12, ΔMFCC12, ΔΔMFCC12,
      Loudness, ΔLoudness, ΔΔLoudness,
      F0, VoiceProb, ZCR, HNR, ΔF0, ΔVoiceProb, ΔZCR, ΔHNR
  - **Teacher:** ① Hard-target
    ② Soft-target [Fayek+, 16]
    } **baseline**
    ③ Modified soft-target
  - **Train data:** *clear / ambiguous / clear + ambiguous*
- ✓ **Evaluation measures**
  - Weighted Accuracy (WA): overall accuracy
  - Unweighted Accuracy (UA): average recall of emotion classes
    - ➢ Average results of 5 trials of training

NTT

Results

**Moderate performance with *ambiguous* data alone,
and best with *clear* + *ambiguous* data**

	Teacher	Train set: *clear*	Train set: *ambig.*	Accuracy [%]: WA	Accuracy [%]: UA
MajorityClass (All *Neutral*)				40.8	25.0
Baseline	hard-target	✓		58.6	53.7
	soft-target	✓		58.1	54.9
Proposed	Modified soft-target	✓		58.5	57.4
			✓	53.6	54.0
		✓	✓	**62.6**	**63.7**

Overall Acc.

Avg. Recall

Results

Moderate performance with *ambiguous* data alone, and best with *clear* + *ambiguous* data

	Teacher	Train set *clear*	Train set *ambig.*	Accuracy [%] WA	Accuracy [%] UA
MajorityClass (All *Neutral*)				40.8	25.0
Baseline	hard-target	✓		58.6	53.7
	soft-target	✓		58.1	54.9
Proposed	Modified soft-target	✓		58.5	57.4
			✓	53.6	54.0
		✓	✓	**62.6**	**63.7**

Moderate performance
even they have been ignored for training!

Results

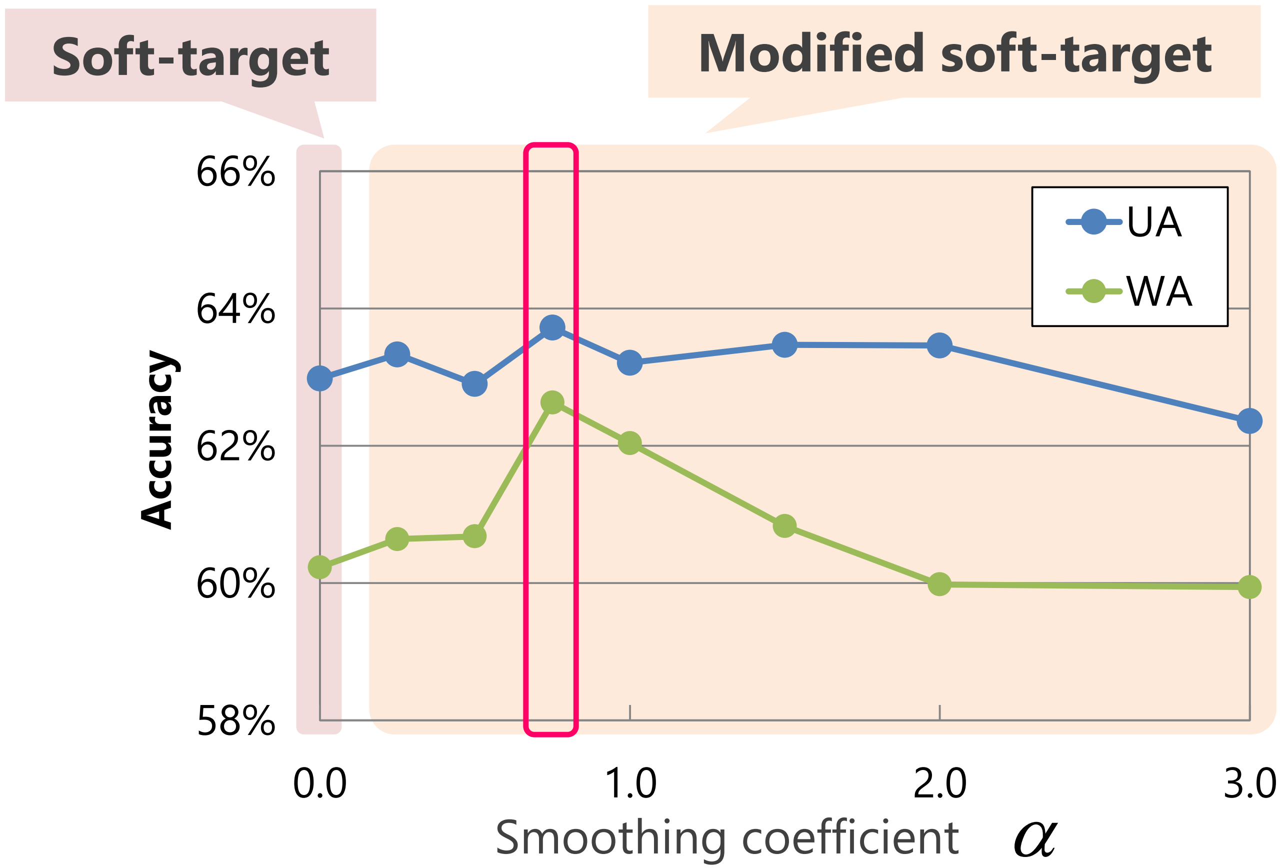
Moderate performance with *ambiguous* data alone, and best with *clear* + *ambiguous* data

	Teacher	Train set: *clear*	Train set: *ambig.*	Accuracy [%]: WA	Accuracy [%]: UA
MajorityClass (All *Neutral*)				40.8	25.0
Baseline	hard-target	✓		58.6	53.7
	soft-target	✓		58.1	54.9
Proposed	Modified soft-target	✓		58.5	57.4
			✓	53.6	54.0
		✓	✓	**62.6**	**63.7**

Best performance

Comparisons of teacher labels

Modified soft-target with smoothing coeff. = 0.75
is better than (conventional) soft-target



Setup

Train: *clear + ambig.*

Model: BLSTM-att

Conclusions

- **Summary**
  - **Purpose:** emotion classification from acoustic features
  - **Approach:** Utilizing *ambiguous* emotional utterances
    to mitigate training data limitation problem
  - **Method:** Soft-target training which deals both
    clear and *ambiguous* emotional utterances in same criteria
    - Equal to ML/MAP estimation of true emotion distributions
  - **Results:** Performances were improved (WA 58.6→62.6%)
    Show the effectiveness of *ambiguous* data for training

- **Future works**
  - Evaluations by other corpus / emotion set
  - Improve modified soft-target (prior distribution of MAP estimation)