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
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%
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

![Diagram showing frame-wise acoustic features and BLSTM-RNNs](image)

- Posterior of emotions
- LSTM Classifier (BLSTM-attention)
- Frame-wise features (F0, MFCC, etc)
- Utterance
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

Issue  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 annotations

Train / Test data

Ground Truth

Happy

Happy

Neutral

Happy

Others (excited)

Others (excited)

Happy

Neutral

Angry

No use

Ground Truth
Approach (1/2)

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

Target emotions

*Neutral, Happy, Sad, Angry*

Clear emo. utter.
Target emo. is dominant

- [Happy, Happy, Happy]
- [Happy, Happy, Neutral]

Ambiguous emo. utter.
Target emo. is minor

- [Happy, Neutral, Angry]
- [Happy, Others, Others]

Not included
- [Others, Others, Others]

Annotation example

- No dominant
- Non-target is dominant
Utilize *ambiguous emotional utterances* (target emo. are minor) to mitigate training data limitation

**Target emotions**
Neutral, Happy, Sad, Angry

**Conventional training**
- **Clear emo. utter.**
  - Target emo. is dominant
  - Annotation example
    - [Happy, Happy, Happy]
    - [Happy, Happy, Neutral]

- **Ambiguous emo. utter.**
  - Target emo. is minor
  - Annotation example
    - [Happy, Neutral, Angry]
    - [Happy, Others, Others]

- **Not included**
  - Annotation example
    - [Others, Others, Others]
Utilize *ambiguous emotional utterances* (target emo. are minor) to mitigate training data limitation.

**Target emotions**
Neutral, Happy, Sad, Angry

**Annotation example**
- [Happy, Happy, Happy]
- [Happy, Happy, Neutral]
- [Happy, Neutral, Angry]
- [Happy, Others, Others]

**Conventional training**
- *Clear emo. utter.*
  Target emo. is dominant

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

**Not included**
Are there no *Happy* characteristics?
### Approach (1/2)

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

**Target emotions**
- Neutral, Happy, Sad, Angry

**Proposed training**

- **Clear emo. utter.**
  - Target emo. is dominant
  - [Happy, Happy, Happy]
  - [Happy, Happy, Neutral]

- **Ambiguous emo. utter.**
  - Target emo. is minor
  - [Happy, Neutral, Angry]
  - [Happy, Others, Others]

- **Not included**
  - [Others, Others, Others]
Control discriminativity to handle both clear and ambiguous emotional utterances

- **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]
    - [Others, Others, Others]

- **Not included**
Proposed

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

✓ Two types of soft-target

1. Soft-target [Fayek+, 16]

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

2. Modified soft-target

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

✓ Model parameters are updated by cross-entropy loss

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

Annotation frequency (sum=1)

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

Additive smoothed form of conventional soft-target

- $\alpha$: Smoothing coefficient
Proposed: modified soft-target

Modified soft-target is suitable to represent ambiguous emotional utterances

✓ Examples of teachers $q(c_k)$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[Happy, Happy, Happy]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Hap, 1.0]</td>
<td>[Hap, 1.0]</td>
<td>[Hap, 0.14; Sad, 0.58; Neu, 0.14; Ang, 0.14]</td>
</tr>
<tr>
<td></td>
<td>Neu, 0</td>
<td>Neu, 0</td>
<td>Neu, 0.14</td>
</tr>
<tr>
<td></td>
<td>Hap, 0</td>
<td>Hap, 0</td>
<td>Sad, 0.58</td>
</tr>
<tr>
<td></td>
<td>Sad, 0</td>
<td>Sad, 0</td>
<td>Neu, 0.14</td>
</tr>
<tr>
<td></td>
<td>Ang, 0</td>
<td>Ang, 0</td>
<td>Ang, 0.14</td>
</tr>
<tr>
<td>[Happy, Happy, Neutral]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Hap, 1.0]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neu, 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hap, 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sad, 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ang, 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Happy, Others, Others]</td>
<td></td>
<td>(no use)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Smoothing coeff. $\alpha = 1$)
Proposed: modified soft-target

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

✓ Examples of teachers $q(c_k)$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[Happy, Happy, Happy]</td>
<td><img src="#" alt="Bar chart" /></td>
<td><img src="#" alt="Bar chart" /></td>
<td><img src="#" alt="Bar chart" /></td>
</tr>
<tr>
<td>[Happy, Happy, Neutral]</td>
<td><img src="#" alt="Bar chart" /></td>
<td><img src="#" alt="Bar chart" /></td>
<td><img src="#" alt="Bar chart" /></td>
</tr>
<tr>
<td>[Happy, Others, Others]</td>
<td>(no use)</td>
<td><img src="#" alt="Bar chart" /></td>
<td><img src="#" alt="Bar chart" /></td>
</tr>
</tbody>
</table>

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 \( q(c_k) \)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[Happy, Happy, Happy]</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>[Happy, Happy, Neutral]</td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>[Happy, Others, Others]</td>
<td>(no use)</td>
<td></td>
<td><img src="image7" alt="Graph" /></td>
</tr>
</tbody>
</table>

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 $q(c_k)$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[Happy, Happy, Happy]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Happy, Happy, Neutral]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Happy, Others, Others]</td>
<td>(no use)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lower discriminativity in ambiguous emo. uttr.
Interpretation

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

Utterance

Annotations

Objective function of the model

“true” distribution of target emo.

Sampling
(N=# of annotations)

[Happy, Happy, Sad]

Objective function of the model

Hap Sad Neu Ang

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

Utterance

Annotations

Objective function of the model

**Discrimination rule (0/1)**

<table>
<thead>
<tr>
<th>Neu</th>
<th>Hap</th>
<th>Sad</th>
<th>Ang</th>
</tr>
</thead>
</table>

“true” distribution of target emo.

Sampling (N=# of annotations)

[Happy, Happy, Sad]
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

Discrimination rule (0/1)

Hard-target

Soft-target

ML-based distribution

MAP-based distribution

Modified soft-target
Interpretation

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

Utterance

Annotations

Objective function of the model

Sampling (N=# of annotations)

“true” distribution of target emo.

[Happy, Happy, Sad]

Discrimination rule (0/1)

Hard-target

ML-based distribution

Soft-target

MAP-based distribution

Modified soft-target

Uniform prior

Objective function of the model

Hap Sad Neu Ang

Hap Sad Neu Ang

Hap Sad Neu Ang

Hap Sad Neu Ang
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

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Neutral</th>
<th>Happy</th>
<th>Sad</th>
<th>Angry</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clear</td>
<td>3548</td>
<td>1324</td>
<td>460</td>
<td>890</td>
<td>874</td>
<td>-</td>
</tr>
<tr>
<td>ambiguous</td>
<td>3693</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3693</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>942</td>
<td>384</td>
<td>135</td>
<td>194</td>
<td>229</td>
<td>-</td>
</tr>
</tbody>
</table>
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]
    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
## Results

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

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Train set</th>
<th>Accuracy [%]</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clear</td>
<td>ambig.</td>
<td>WA</td>
<td>UA</td>
</tr>
<tr>
<td>MajorityClass</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(All Neutral)</td>
<td></td>
<td></td>
<td>40.8</td>
<td>25.0</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard-target</td>
<td>✔</td>
<td></td>
<td>58.6</td>
<td>53.7</td>
</tr>
<tr>
<td>Soft-target</td>
<td>✔</td>
<td></td>
<td>58.1</td>
<td>54.9</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modified soft-target</td>
<td>✔</td>
<td></td>
<td>58.5</td>
<td>57.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Acc.</td>
<td>✔</td>
<td>✔</td>
<td>62.6</td>
<td>63.7</td>
</tr>
</tbody>
</table>

Moderate performance with ambiguous data alone, and best with clear + ambiguous data.
# Results

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

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Train set</th>
<th>Accuracy [%]</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clear</td>
<td>ambig.</td>
<td>WA</td>
<td>UA</td>
</tr>
<tr>
<td>MajorityClass (All Neutral)</td>
<td></td>
<td></td>
<td>40.8</td>
<td>25.0</td>
</tr>
<tr>
<td>Baseline</td>
<td>Hard-target</td>
<td>✔</td>
<td>58.6</td>
<td>53.7</td>
</tr>
<tr>
<td></td>
<td>Soft-target</td>
<td>✔</td>
<td>58.1</td>
<td>54.9</td>
</tr>
<tr>
<td>Proposed</td>
<td>Modified soft-target</td>
<td>✔</td>
<td>58.5</td>
<td>57.4</td>
</tr>
<tr>
<td></td>
<td>Soft-target</td>
<td>✔</td>
<td>53.6</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td>✔</td>
<td>✔</td>
<td>62.6</td>
<td>63.7</td>
</tr>
</tbody>
</table>

*Moderate performance even they have been ignored for training!*
## Results

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

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Train set</th>
<th>Accuracy [%]</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>clear</td>
<td>ambig.</td>
<td>WA</td>
</tr>
<tr>
<td>MajorityClass (All <em>Neutral</em>)</td>
<td></td>
<td>40.8</td>
<td>25.0</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>Hard-target ✔</td>
<td>58.6</td>
<td>53.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Soft-target ✔</td>
<td>58.1</td>
<td>54.9</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>Modified soft-target ✔</td>
<td>58.5</td>
<td>57.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>53.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>62.6</td>
</tr>
</tbody>
</table>

Best performance
Comparisons of teacher labels

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

Soft-target

Modified soft-target

Accuracy

UA

WA

Smoothing coefficient \( \alpha \)

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)