ICASSP
Calgary, Canada, 19 April 2018

Maximal Figure-of-Merit Embedding for Multi-label Audio Classification

Joined work by Ivan Kukanov, Ville Hautamaki, Kong Aik Lee

Presenter: Ivan Kukanov

ivan@kukanov.com

A*STAR Institute for Infocomm Research

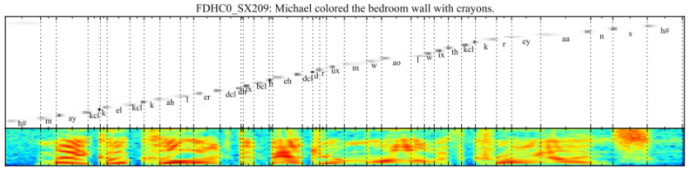
UNIVERSITY OF EASTERN FINLAND

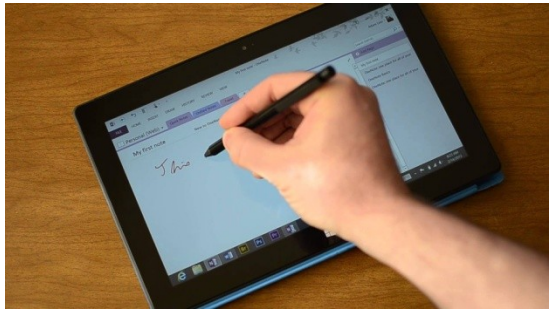
Problem Statement

Single-label

- Speech recognition
- Handwriting recognition
- Language recognition

...





Multi-label

- Image tagging
- Acoustic scene recognition
- Functional genomics: predicting the gene functional classes

...

Comparison: None

Indication: Burmese male has complete TB treatment

Findings: Both lungs are clear and expanded with no infiltrates. Basilar focal atelectasis is present in the lingula. Heart size normal. Calcified right hilar XXXX are present

Impression: No active disease.

MeSH

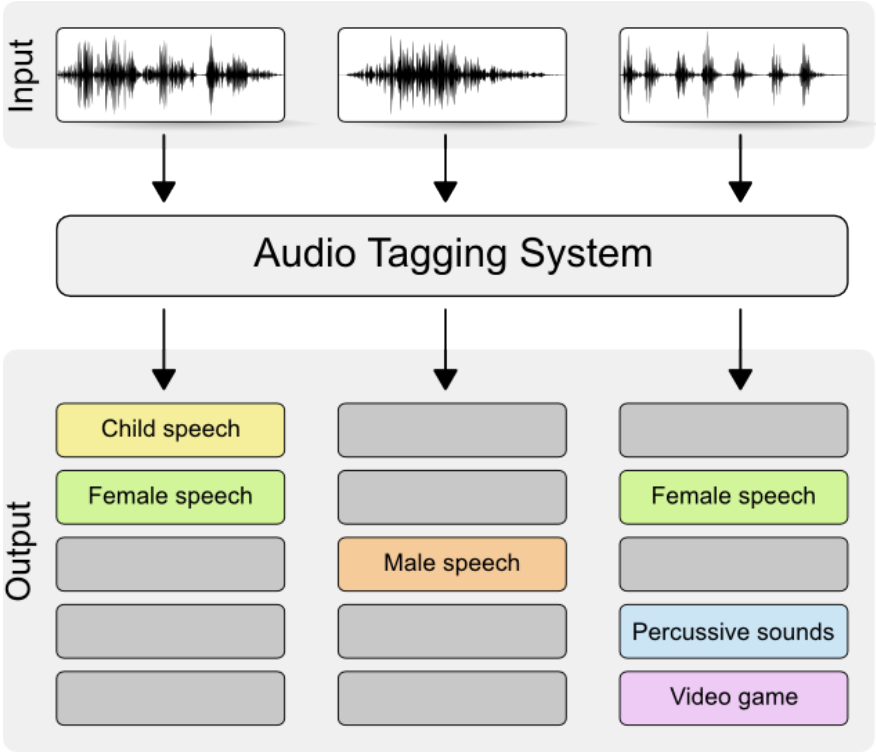
Major

Pulmonary Atelectasis / lingula / focal
Calcinosis / lung / hilum / right

CHEST 2V FRONTAL/LATERAL XXXX, XXXX XXXX PM

Problem Statement

- Multi-label acoustic event detection
  - Find most relevant subset of labels



Ref.: [1]

Problem Statement

- Multi-label acoustic event detection
  - Training data

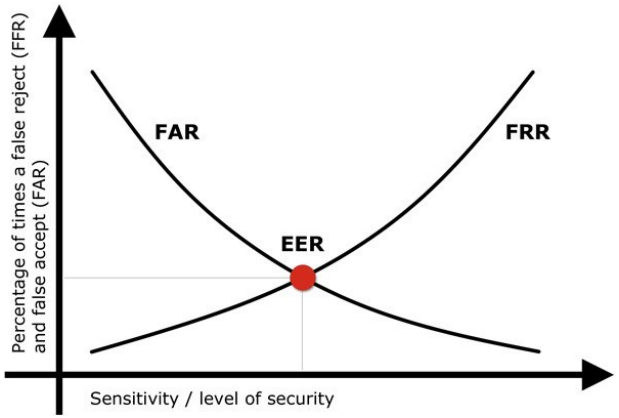
$$\mathbb{T} = \left\{ (\mathbf{X}_i, \mathbf{y}_i) | \, i = \overline{1, N} \right\} \qquad \mathbf{X}_i \in \mathbb{R}^D \qquad \mathbf{y}_i \in \{0,1\}^M$$

  - Learn a mapping

$$\mathbf{G} : \mathbb{X} \rightarrow \mathbb{Y}$$

  - Performance measure:

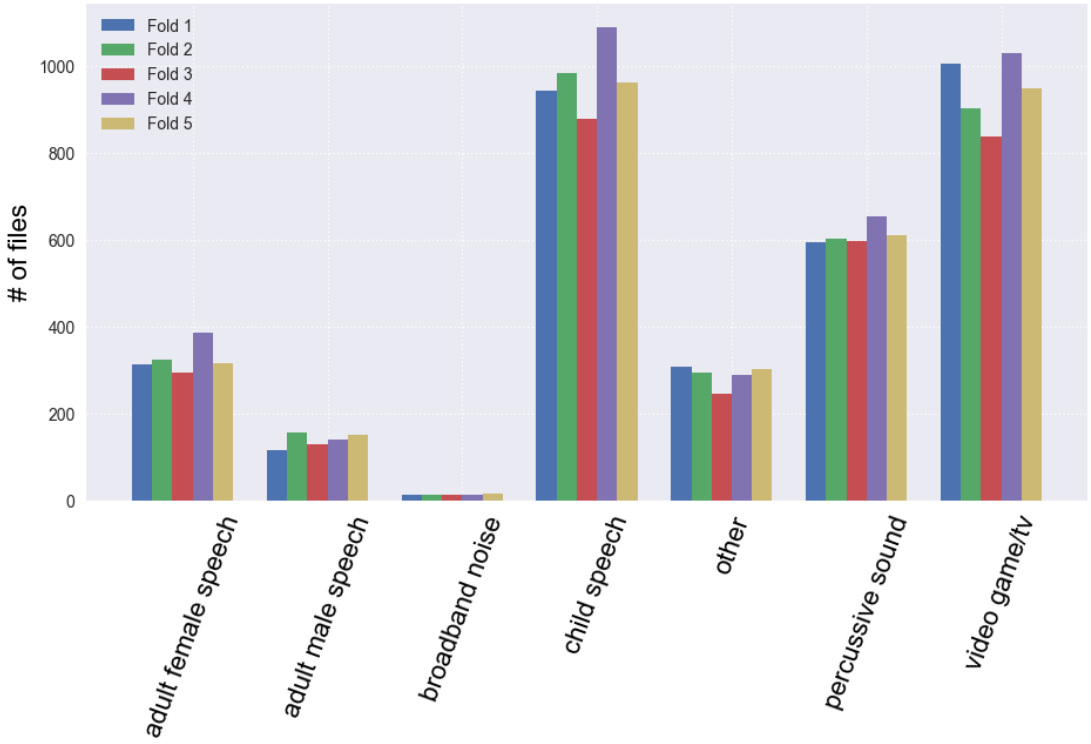
    equal error rate (EER)



Note: cardinality $|\mathbb{Y}| = 2^M, \quad M = 7$

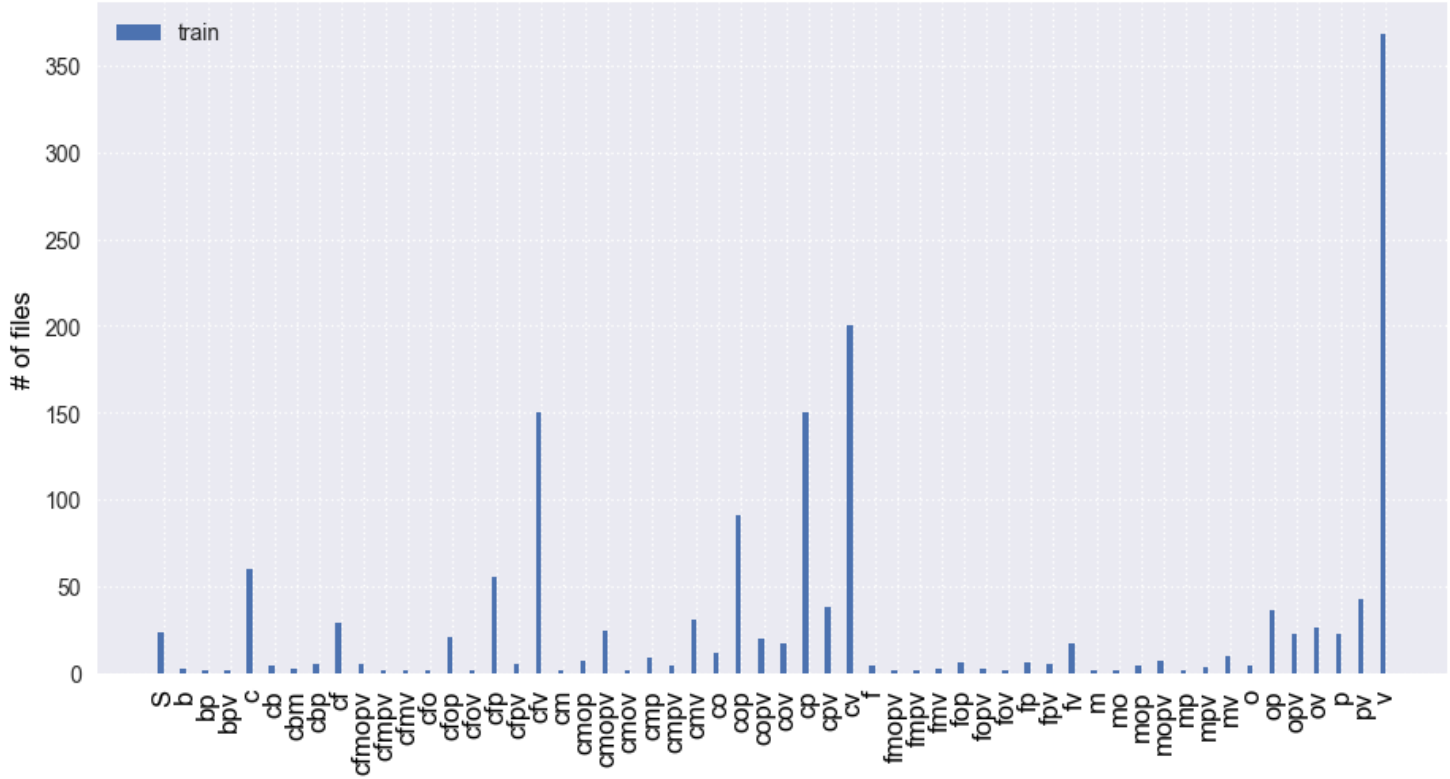
Dataset

- DCASE 2016 challenge: *CHiME-Home* dataset [2]



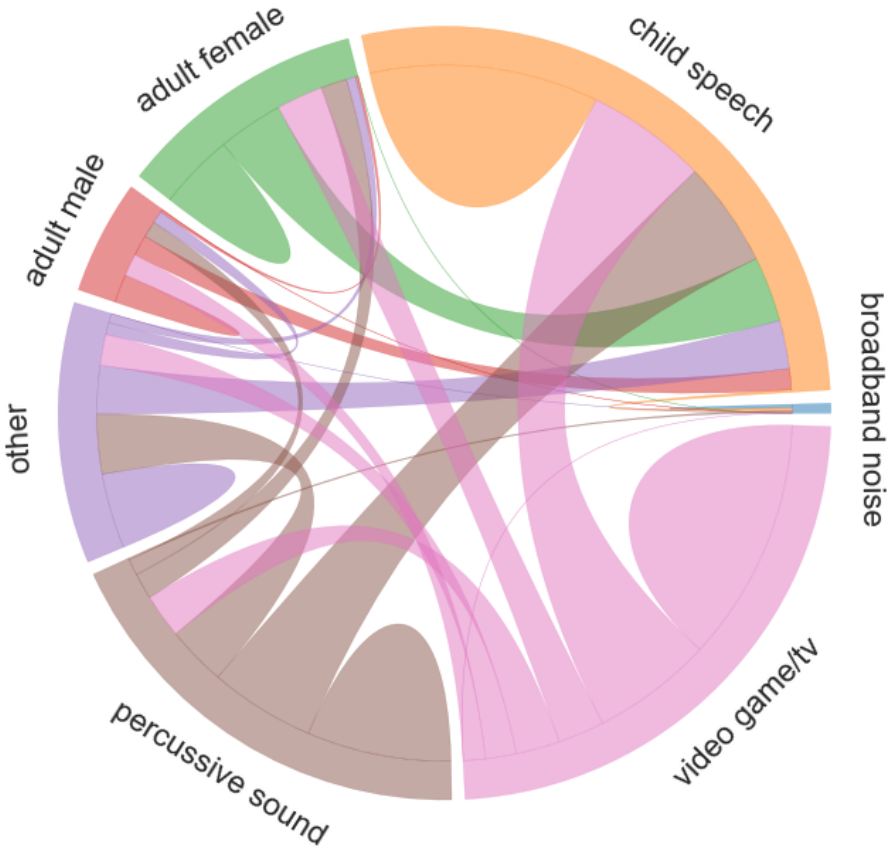
Dataset

- Labels distribution: 56 unique labels



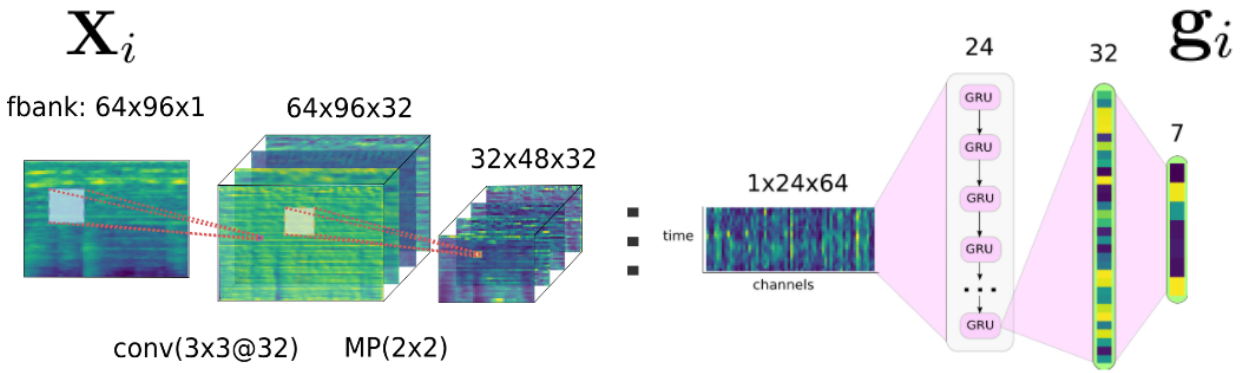
Dataset

- Inter-connection of labels



Baseline Method

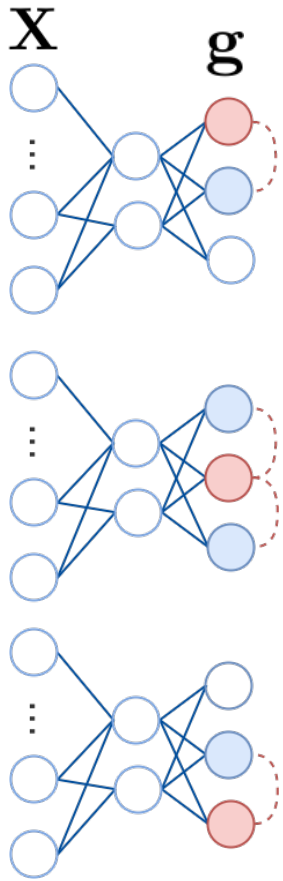
- Model: CNN-RNN
- Output: sigmoid scores
- Objective function: binary cross-entropy (BCE)
- Problem
  - Ignores labels dependence
  - Indirect optimization of the performance measure (EER)



$$J_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left\{\mathbf{y}_i^{\top}\log\left(\mathbf{g}_i\right) + \left(1-\mathbf{y}_i\right)^{\top}\log\left(1-\mathbf{g}_i\right)\right\}$$

MFoM: intuition & motivation

- **Example**: label dependence as *"units-vs-zeros"*, e.g. $\mathbf{y}^{\mathrm{T}}=[1, 0, 1]$



$$\psi_1 = -g_1 + g_2$$

$$\psi_2 = -g_2 + \ln\left\{\frac{1}{2}\left(e^{g_1} + e^{g_3}\right)\right\}$$

$$\psi_3 = -g_3 + g_2$$

- target
- confusing

MFoM: intuition & motivation

1) Misclassification measure: *"units-vs-zeros"*

Target

Confusing

$$\psi_k = -g_k + \ln\left(\frac{1}{|\mathbf{I}|}\sum_{j\in\mathbf{I}} e^{g_j}\right), \quad \text{where} \quad \begin{cases} \text{if } y_k = 1 \Rightarrow \mathbf{I} = \mathbf{y}_{\{0\}} \\ \text{if } y_k = 0 \Rightarrow \mathbf{I} = \mathbf{y}_{\{1\}} \end{cases}$$

k – index of class, $\mathbf{I}$ – set of indices

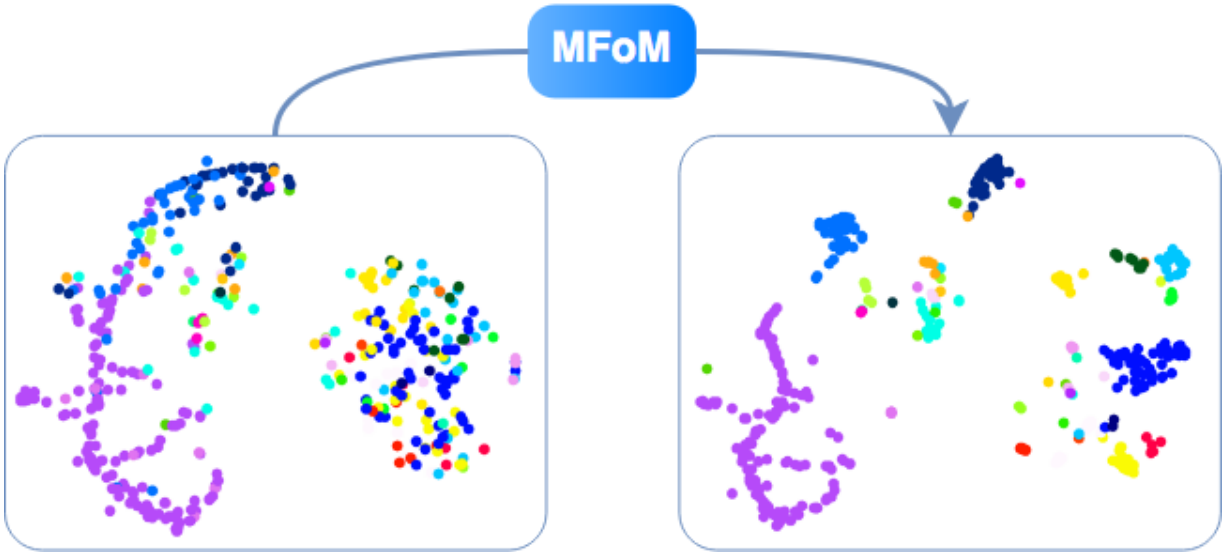
2) Smooth error function (*l - scores*)

$$l_k = \frac{1}{1+\exp\left[-\alpha_k\psi_k - \beta_k\right]}$$

α_k, β_k – scale and shift parameters

MFoM: intuition & motivation

- **Example**: MFoM-transformation, i.e. $\boldsymbol{l}$ - scores

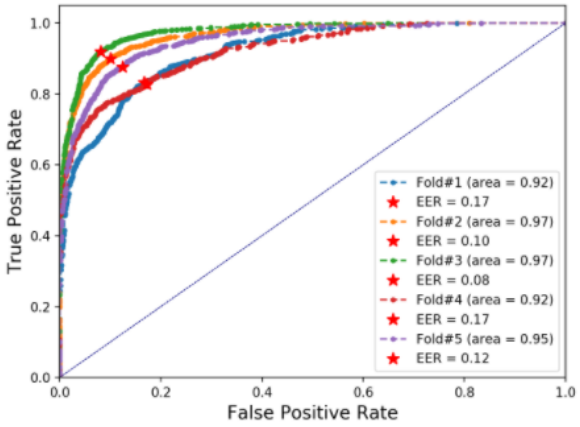


Sigmoid scores, $\boldsymbol{g}$

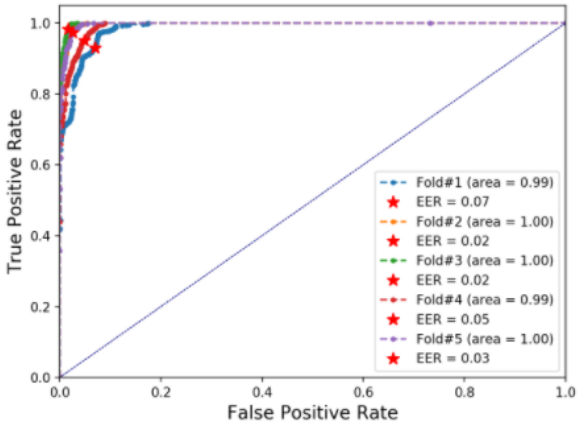
$$\boldsymbol{l} = \mathrm{MFoM}(\mathbf{g}, \mathbf{y})$$

MFoM: intuition & motivation

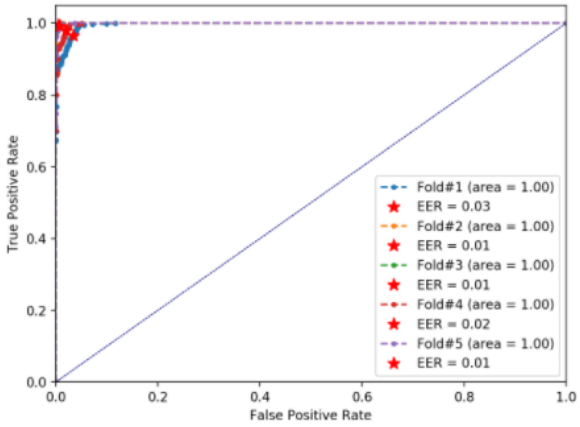
- MFoM-transformation is the ***contraction mapping***, w.r.t. EER metric



Sigmoid scores, $\boldsymbol{g}$



$$\boldsymbol{l}^1 = \mathrm{MFoM}(\mathbf{g}, \mathbf{y})$$

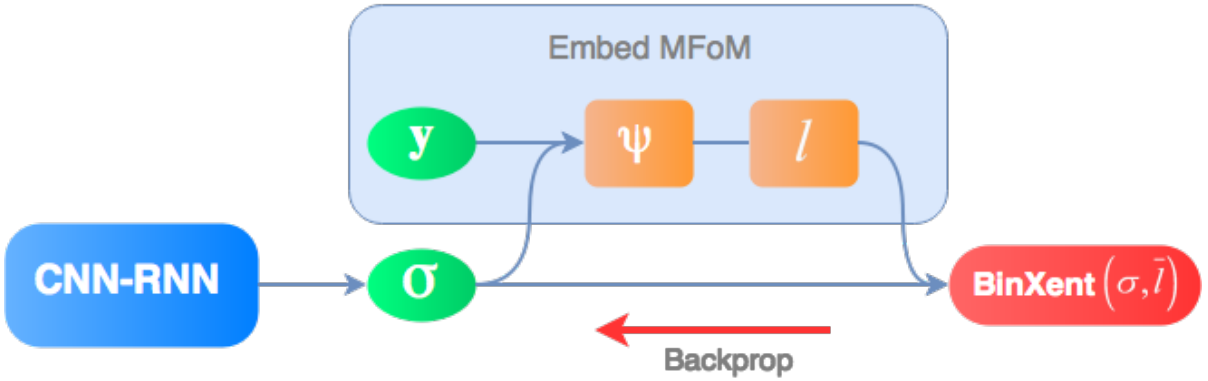


$$\boldsymbol{l}^2 = \mathrm{MFoM}(\boldsymbol{l}^1, \mathbf{y})$$

- Use $\boldsymbol{l}$ - scores as the "*soft labels*"

Proposed Method: MFoM-embedding

- Model: CNN-RNN
- Output: sigmoid scores
- Objective function: binary cross-entropy with MFoM
- *Soft-labels $\boldsymbol{l}$ - scores*

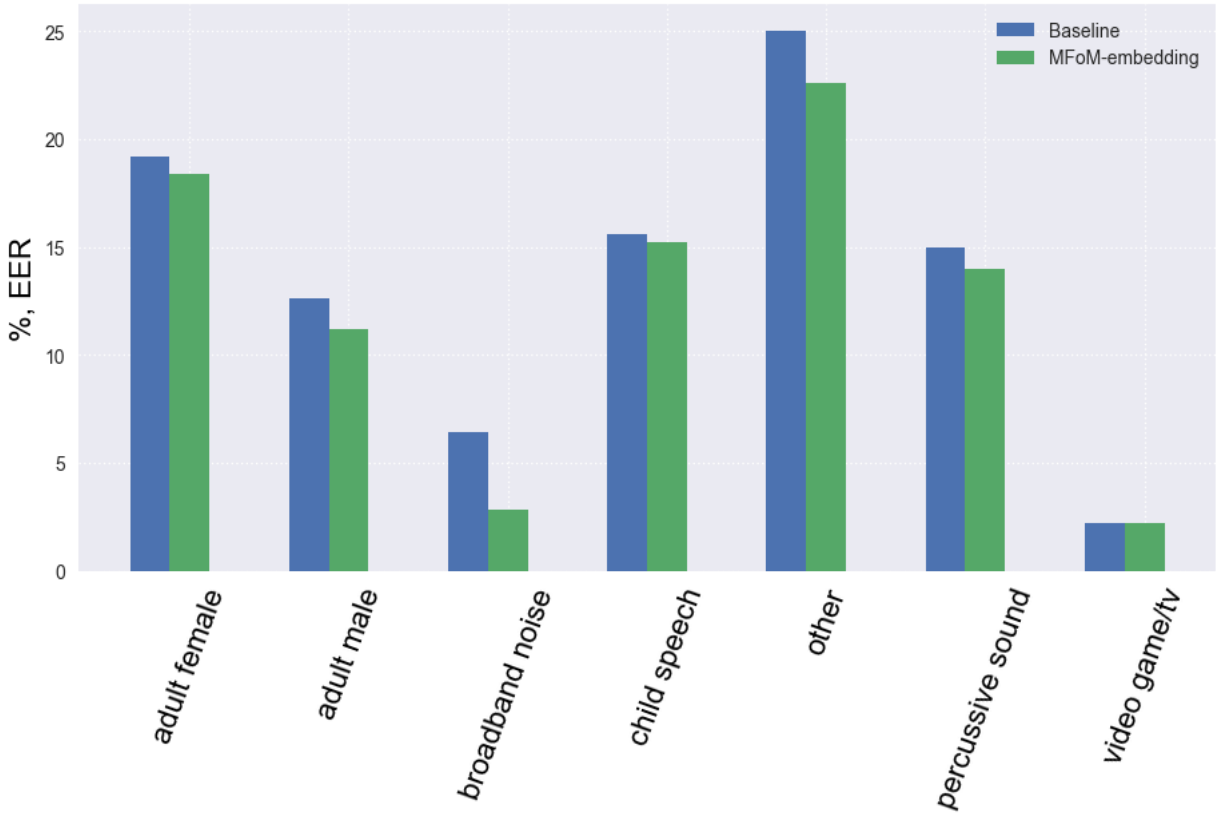


$$J_{\mathrm{MFoM}} = -\frac{1}{N}\sum_{i=1}^{N}\left\{\bar{\boldsymbol{l}}_i^{\top}\log\left(\mathbf{g}_i\right)+\left(1-\bar{\boldsymbol{l}}_i\right)^{\top}\log\left(1-\mathbf{g}_i\right)\right\}$$

$$\bar{\boldsymbol{l}}_i = 1 - \boldsymbol{l}_i$$

Results

- More than 9% relative improvement



Results

- Comparison with the other models. Metric is AvgEER, %

Fold #	GMM [6]	CNN [23]	CRNN [9]	J_{BCE}	J_{MFoM}
1	24.2	.	.	16.0	**15.3**
2	17.1	.	.	11.4	**10.7**
3	17.7	.	.	9.3	**7.9**
4	20.2	.	.	13.9	**13.8**
5	25.3	.	.	18.0	**14.4**
Avg.	20.9	16.6	13.0	13.6	**12.4**

Conclusion

- Applied MFoM transformation to optimize EER metric
- Embed MFoM into DNN objective function
- Instead of using "*hard*" (0/1) labels, used "*soft*" MFoM labels

Thank you for your attention!

Ivan Kukanov

ivan@kukanov.com
www.kukanov.com

Take Away

- Optimization tricks: use Adadelta optimizer
- Scale and shift (α_k, β_k) parameters optimize as in BatchNorm
- Properties of MFoM-transformation
  - a.k.a. soft-labels (*Dark-knowledge* [3])
  - Contraction mapping, w.r.t. EER metric
  - t-SNE and AUC proof

References

[1] DCASE 2016 Challenge: http://www.cs.tut.fi/sgn/arg/dcase2016/task-audio-tagging

[2] Foster et al., "CHiME-Home: A dataset for sound source recognition in a domestic environment", Proc WASPAA, Oct 2015.

[3] G. Hinton et al., “Distilling the Knowledge in a Neural Network”, Mar 2015