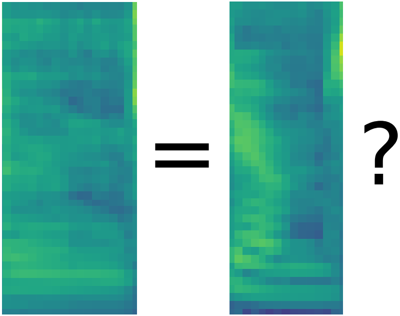
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Benjamin Milde, Chris Biemann

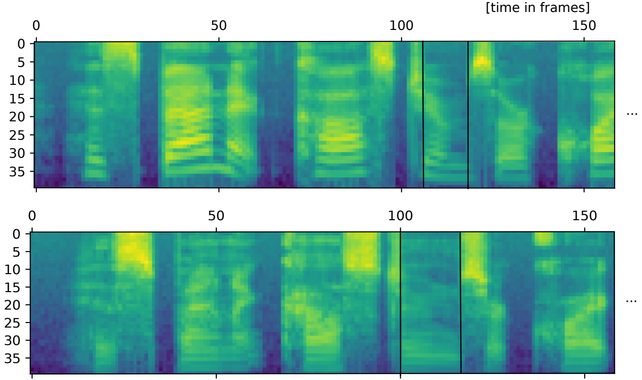
UNSPEECH: UNSUPERVISED SPEECH CONTEXT EMBEDDINGS

Motivation



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Motivation - Context



1

[1] Example in the style of: Aren Jansen, Samuel Thomas, and Hynek Hermansky. 2013. Weak top-down constraints for unsupervised acoustic model training. In ICASSP, pages 8091–8095.

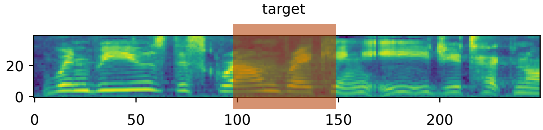
Inspiration - Negative sampling

She had your dark suit in greasy washwater ...

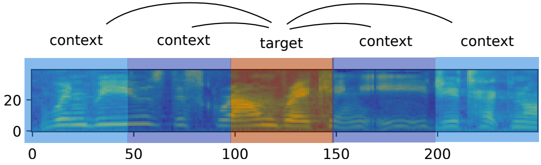
your	dark	suit	in	greasy
context	context	target	context	context

- Word2vec, skipgram with negative sampling
- Binary task instead of directly predicting surrounding words
- Is "dark" + "suit" a context pair?

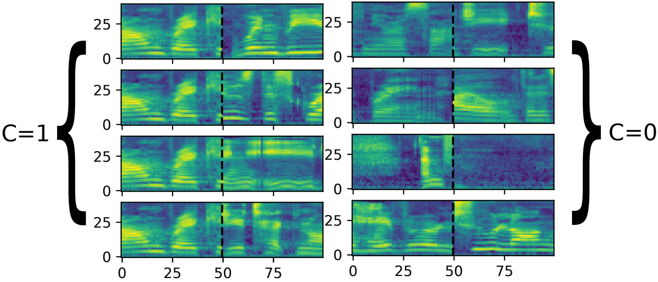
Context example



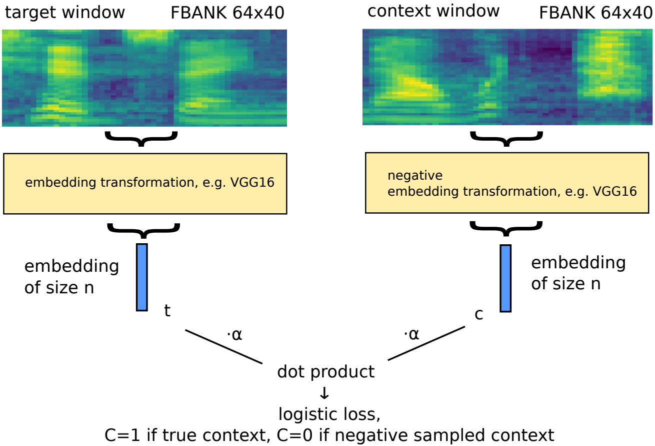
Context example



Example samples



Proposed model



Negative sampling loss

$$NEG_{loss} = -k \cdot log(\sigma(emb_t^T emb_c)) - \sum_{i=1}^{k} log(1 - \sigma(emb_{neg1_i}^T emb_{neg2_i})) \quad (1)$$

- The objective function is similar to negative sampling in word2vec
- But we are not contrasting *emb_t* with a *emb_neg* and choose two random unrelated samples instead for the negative sum.

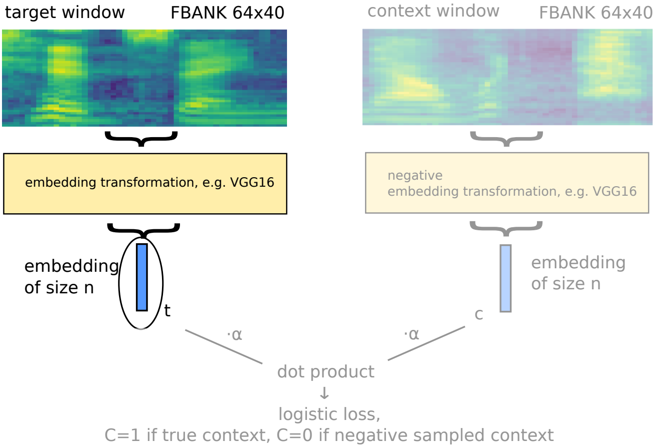
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Negative sampling loss

$$
\begin{aligned}
NEG_{loss} &= -k \cdot log(\sigma(emb_t^T emb_c)) \\
&- \sum_{i=1}^{k} log(1 - \sigma(\underline{emb_{neg1_i}^T emb_{neg2_i}}))
\end{aligned}
\tag{2}
$$

- The objective function is similar to negative sampling in word2vec
- But we are not contrasting *emb_t* with a *emb_neg* and choose two random unrelated samples instead for the negative sum.

Applying a trained unspeech model



Unspeech rep. of an utterance

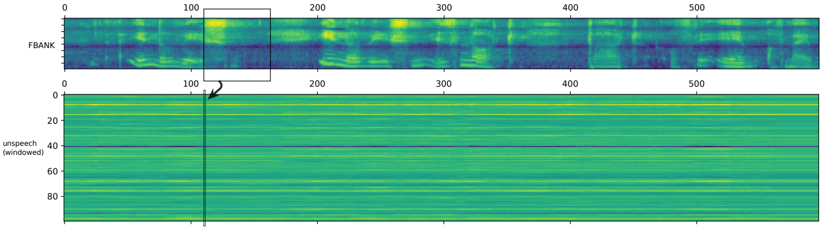


Figure: Windowed unspeech-64 representation

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

TSNE plot TED-LIUM dev set

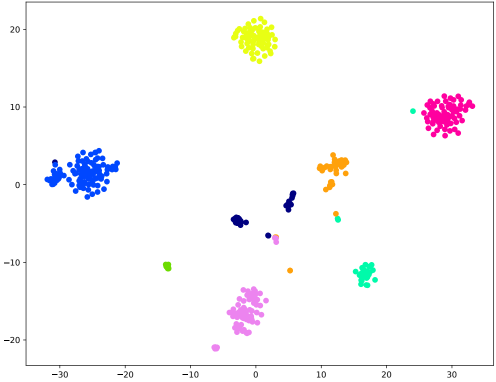
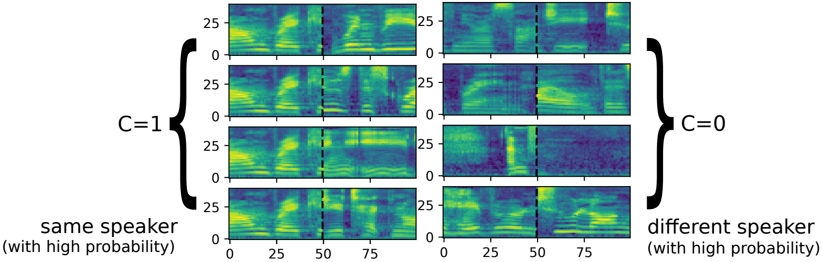


Figure: TSNE plot of unspeech vectors averaged across utterances, TED-LIUM dev set

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Example Samples



Evaluation

- Speaker embedding
- Context clustering
- ASR evaluations with Kaldi:
  - Context clustering $\rightarrow$ cluster-ids in speaker adaptation
  - Providing TDNN-HMM acoustic models with unspeech context embeddings

Evaluation: datasets

Table: Comparison of English speech data sets used in our evaluations

	hours			speakers		
dataset	**train**	**dev**	**test**	**train**	**dev**	**test**
TED-LIUM V2	211	2	1	1273+3	14+4	13+2
Common Voice V1	242	5	5	16677	2728	2768
TEDx (crawled)	9505			41520 talks		

Same/different speaker experiment

Table: Equal error rates (EER) on TED-LIUM V2 – Unspeech embeddings correlate with speaker embeddings.

Embedding	**EER**		
TED-LIUM:	**train**	**dev**	**test**
(1) i-vector	7.59%	**0.46%**	1.09%
(2) i-vector-sp	**7.57%**	0.47%	**0.93%**
(3) unspeech-32-sp	13.84%	5.56%	3.73%
(4) unspeech-64	15.42%	5.35%	2.40%
(5) unspeech-64-sp	13.92%	3.4%	3.31%
(6) unspeech-64-tedx	19.56%	7.96%	4.96%
(7) unspeech-128-tedx	20.32%	5.56%	5.45%

EER = equal error rate, point on a false positive / false negative curve, where both error rates are equal -32 = 32 input frames, -64 = 64 input frames, ...

Context clustering

- Averaged unspeech vectors across time: one 100d vector per utterance
- We use HDBSCAN* to cluster, a modern density based cluster algorithm [2]
- Scales well to large quantities (average case complexity $\approx$ N log N)
- Parameters are easy to set, no epsilon like in vanilla DBSCAN

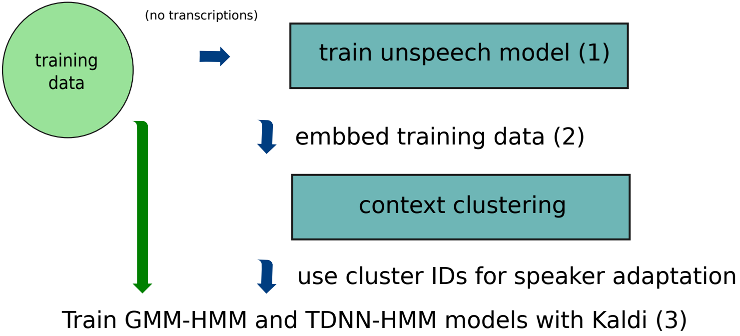
[2] L. McInnes, J. Healy, and S. Astels, HDBSCAN*: Hierarchical density based clustering," The Journal of Open Source Software, vol. 2, no. 11, p. 205, 2017.

Context clustering - NMI

Table: Comparing Normalized Mutual Information (NMI) on clustered utterances from TED-LIUM using i-vectors and (normalized) Unspeech embeddings with speaker labels from the corpus. ”-sp” denotes embeddings trained with speed-perturbed training data.

Embedding	**Num. clusters**			**Outliers**			**NMI**		
	train	**dev**	**test**	**train**	**dev**	**test**	**train**	**dev**	**test**
TED-LIUM IDs	1273 (1492)	14	13	3	4	2	1.0	1.0	1.0
i-vector	1630	12	10	8699	1	2	0.9605	**0.9804**	**0.9598**
i-vector-sp	1623	12	10	9068	1	2	0.9592	**0.9804**	**0.9598**
unspeech-32-sp	1686	16	12	3235	22	32	**0.9780**	0.9536	0.9146
unspeech-64	1690	16	11	5690	14	21	0.9636	0.9636	0.9493
unspeech-64-sp	1702	15	11	3705	23	25	0.9730	0.9633	0.9366

Context clustering for ASR



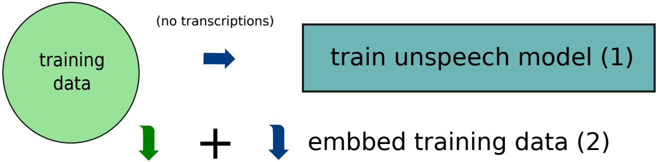
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Context clustering for ASR
WER results

Table: WERs for different context IDs for speaker adaptation in TDNN-HMM ASR models. (One speaker per talk, one speaker per utterance, unspeech hdbscan IDs)

Acoustic model	Spk. div.	Dev WER	Test WER
GMM-HMM	per talk	18.2	16.7
TDNN-HMM		7.8	8.2
GMM-HMM	per utt.	18.7	19.2
TDNN-HMM		7.9	9.0
GMM-HMM	Unspeech 64	17.4	**16.5**
TDNN-HMM		7.8	**8.1**

Unspeech contexts in TDNN-HMMs



TDNN-HMM models (3)

append context vectors to input

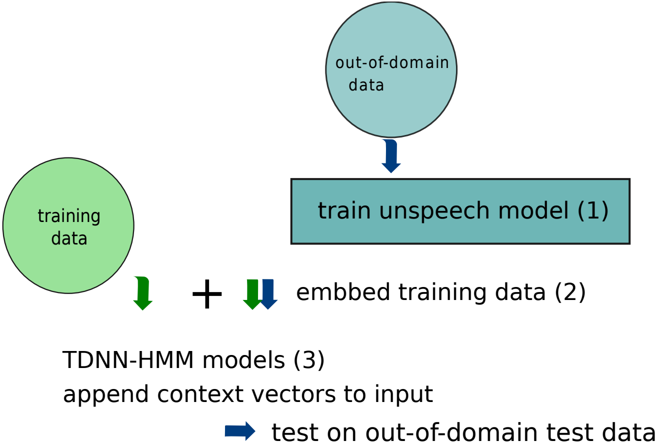
Note that the standard TDNN-HMMs recipes in Kaldi also use ivectors (speaker vectors) similarly.

Unspeech contexts in TDNN-HMMs

Table: WER for TDNN-HMM chain models trained with Unspeech embeddings on TED-LIUM.

Context vector	**Dev WER**	**Test WER**
(1) none	8.5	9.1
(2) i-vector-sp-ted	**7.5**	8.2
(3) unspeech-64-sp-ted	8.3	9.0
(4) unspeech-64-sp-cv	8.3	9.1
(5) unspeech-64-sp-cv + (2)	7.6	**8.1**
(6) unspeech-64-tedx	8.2	8.7
(7) unspeech-128-tedx	8.2	8.9

Unspeech contexts in TDNN-HMMs



Unspeech contexts in TDNN-HMMs WER results out-of-domain test data

Table: Training on TED-LIUM and decoding on Common Voice V1.

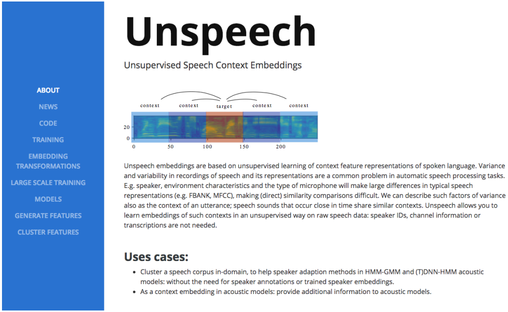
Context vector	**Dev WER**	**Test WER**
(1) none	29.6	28.5
(2) i-vector-sp-ted	29.0	28.2
(3) unspeech-64-sp-cv	**27.9**	**26.9**
(4) unspeech-64-sp-cv + (2)	28.2	27.4
(5) unspeech-64-tedx	28.8	27.5
(6) unspeech-128-tedx	28.7	28.0

Conclusion

- We showed a simple unsupervised context embedding method, that can be trained on large amounts of unlabelled data
- Our context embeddings contain speaker characteristics
- Our method can be used for context clustering
- Context cluster ids can aid speaker adaptation in acoustic models when no speaker information is available
- Can help in domain adaptation, when the unspeech models are trained on unlabelled data of the target domain

http://unspeech.net

- http://unspeech.net - Download model source code (Python3/Tensorflow), pretrained models and documentation

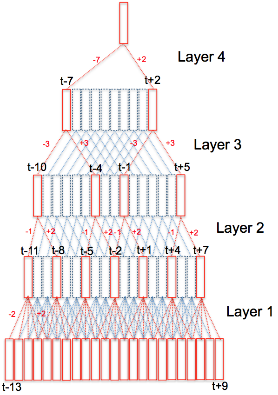


Questions?

- now
- after the session
- or mail me: milde@informatik.uni-hamburg.de

Extra slides

Unspeech contexts in TDNN-HMMs



Stationary hypothesis

Automatic Speaker Clustering (Jin et. al 1997):[3]

- ”Our algorithm takes the advantage [...] that consecutive segments are more likely to come from the same speaker”
- ”In practice, we regard speaker as a generic concept which really means speaker with channel and background condition”
- We call this generic concept ”context”

[3] H. Jin, F. Kubala, and R. Schwartz, ”Automatic speaker clustering,” in Proceedings of the DARPA speech recognition workshop, 1997, pp. 108–111