Towards Unsupervised Subject-Independent Speech-Based Relapse Detection in Patients with Psychosis using Variational Autoencoders

C. Garoufis[1], A. Zlatintsi[1], P. P. Filntisis[1], N. Efthymiou[1], E. Kalisperakis[2,3], T. Karantinos[2], V. Garyfalli[2,3], M. Lazaridi[2,3], N. Smyrnis[2,3], P. Maragos[1]

[1]School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2]Laboratory of Cognitive Neuroscience, University Mental Health Research Institute, Athens, Greece
[3]National & Kapodistrian University of Athens, Medical School

cgaroufis@mail.ntua.gr, {filby,nefthymiou}@central.ntua.gr, {nzlat, maragos}@cs.ntua.gr, smyrnis@med.uoa.gr

Introduction – Problem Definition

What is a relapse?

- Deterioration in the condition of mental patients.
- Signs of relapses appear in various modalities, including **speech**.
  - Bipolar Disorder: longer pauses between utterances, increased formant frequencies.
  - Schizophrenia: lower speech rate, decreased formant frequencies.

- Goal: Being able to **detect** and **predict** the appearance of relapses from spontaneous speech in patients in the psychotic spectrum.
  - Validation of subjective clinician evaluations.
  - Ability to intervene by predicting the appearance of relapses.

- Majority of studies in the literature tackle the problem using supervised learning approaches, either feature-based [1] or using deep learning [2].

[1] J. Gideon, E. M. Provost, and M. McInnis, "Mood State Prediction from Speech of Varying Acoustic Quality for Individuals with Bipolar Disorder," in Proc. ICASSP 2016
[2] L. He and C. Cao, "Automated Depression Analysis using Convolutional Neural Networks from Speech," Journal of Biomedical Informatics, vol. 83, pp. 103–111, 2018.

Motivation - Contribution

Motivation:

- Appearance of relapses in patients is scarce -> opting for an **unsupervised approach** in an anomaly detection framework, since models can be trained only using data from stable time periods.
- Previous work used a deterministic Convolutional Autoencoder [3] for personalized relapse detection and prediction.
- How can we scale this approach at a **universal** (patient-independent) setting?

Contribution:

- Development of a **Convolutional Variational Autoencoder** (CVAE) on data collected from patient – clinician interviews.
- Training at spectrogram level, results aggregated in a per-session basis.
- Personalized models: **Comparable** performance between CVAEs and CAEs.
- Universal models: CVAEs **significantly outperform** CAEs, reach the performance of personalized models in conjunction with **personalized normalization** and norm pooling for temporal aggregation.

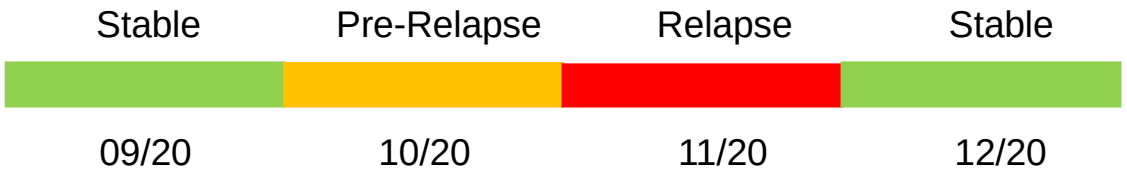
[3] C. Garoufis, A. Zlatintsi, P. Filntisis, N. Efthymiou, E. Kalisperakis, T. Karantinos, V. Garyfalli, L. Mantonakis, N. Smyrnis, and P. Maragos, "An Unsupervised Learning Approach for Detecting Relapses from Spontaneous Speech in Patients with Psychosis," in Proc. BHI 2021

30th EUSIPCO EUROPEAN SIGNAL PROCESSING CONFERENCE
BELGRADE/SERBIA
29 AUG – 2 SEP 2022

Data Collection

- Study participants: 24 patients in the bipolar or psychotic spectrum.
- Annotations of the condition of the patients as stable or relapsing by the expert clinicians, based on:
  - Monthly in-person clinical assessments between the patients and experienced clinicians, through which psychopathological scales are estimated.
  - Weekly unstructured (duration: 5-10 min) interviews conducted between patients and clinicians via a dedicated tablet app and then stored in a cloud server.
  - Communication between the clinicians and the patient's environment.
- In this work, we will use the **short unstructured interviews** between patients and clinicians.

- Further data categorization as:
  - **Clean**: Patient condition is annotated as stable.
  - **Relapse**: A relapse has been detected by the clinicians.
  - **Pre-relapse**: Interviews conducted up to 30 days prior to the appearance of a relapse.



- Both pre-relapse and relapse data are considered as **anomalous**.

30th EUSIPCO
EUROPEAN SIGNAL PROCESSING CONFERENCE
BELGRADE/SERBIA
29 AUG - 2 SEP 2022

Data Preprocessing

Collected Dataset Statistics:

- Total amount of data used: **375 interviews**, from **13 patients**.
- 8 of the patients experienced a relapse during the course of the study.
- The rest were selected on the basis of amount of available speech data.

Patient demographics, as well as statistics on the amount of recorded and analyzed speech utterances.

Demographics	
Male/Female	8/5
Age (years)	27.5 ± 6.7
Education (years)	13.5 ± 1.9
Illness dur. (years)	7.9 ± 7.6
Recorded Data	
Num. of Interviews (total)	375
Num. of Interviews (mean±std)	28.8 ± 8.7
Diarized speech duration (in sec)	30509
Diarized speech duration (in sec, mean±std)	2347 ± 1550
Num. of Utterances (total)	12107
Num. of Utterances (mean±std)	931 ± 527
Num. of Utterances (clean, mean±std)	754 ± 425
Num. of Utterances (pre-relapse, mean±std)	119 ± 126
Num. of Utterances (relapse, mean±std)	169 ± 162

Patient Speech Isolation:

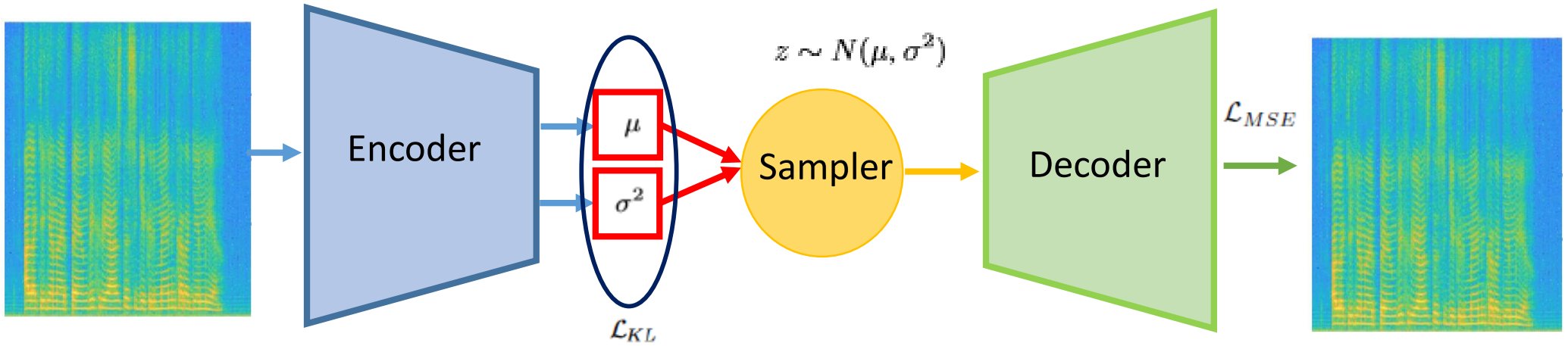
- Audio extracted from interviews, and downsampled to 16 kHz.
- Speech excerpts corresponding to the patients isolated using kaldi.
- Final utterance statistics: **12107 utterances/30509 sec**.

Feature Extraction:

- Computed **mel-spectrograms** for each speech utterance.
- Parameters: 512-sample window, 256-sample hop length, 128 mel bands.
- Spectrograms cut at slices of 64 frames (ca 1 sec.) -> 128x64 input representation, then standardized and log-scaled.

30th EUSIPCO EUROPEAN SIGNAL PROCESSING CONFERENCE
BELGRADE/SERBIA
29 AUG - 2 SEP 2022

Methodology: Variational Autoencoders

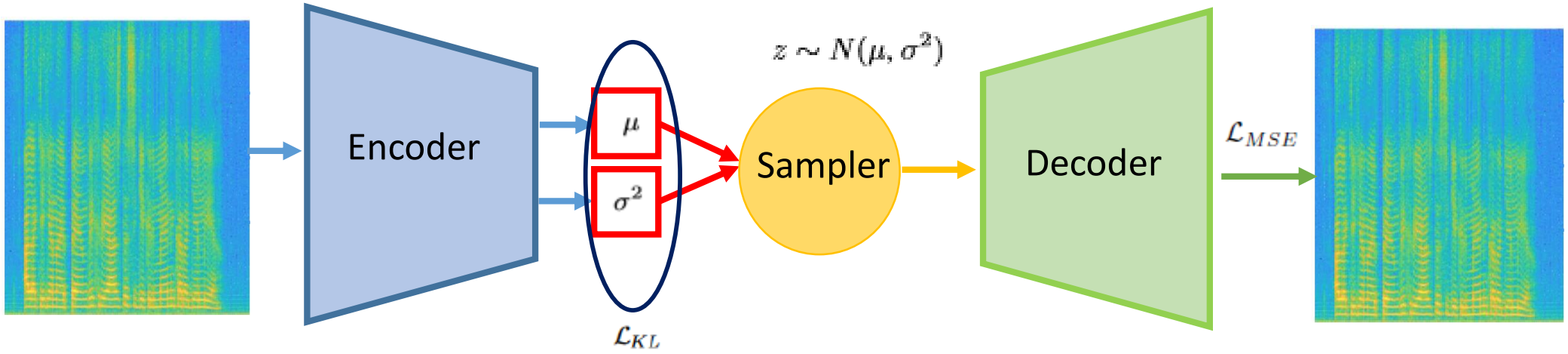


- Probabilistic variant of classical autoencoders, first developed in [4]
- Encoder (inference model): **Encodes** its input into a low-dimensional latent representation, assumed to follow an isotropic Gaussian distribution.
- Decoder (generative model): Attempts to **reconstruct** the input from a sample drawn from the learned distribution.
- Used in a variety of audio-related tasks, such as speech enhancement [5] and speech representation learning [6]

[4] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, 2013.
[5] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder", in Proc. ICASSP 2021
[6] J. Chorowski, R. Weiss, S. Bengio, and A. van den Oord, "Unsupervised Speech Representation Learning using WaveNet Autoencoders," IEEE/ACM Trans. on Audio, Speech, and Language Process., vol. 27, no. 12, pp. 2041–2053, 2019.

30th EUSIPCO EUROPEAN SIGNAL PROCESSING CONFERENCE
BELGRADE/SERBIA
29 AUG – 2 SEP 2022

Methodology: Variational Autoencoders



- **Encoder:** 3 convolutional blocks, alternating 2D-convolutional layers and 2D max pooling layers, **increasing** number of filters + pair of parallel layers to estimate μ and σ^2 .
- **Decoder:** 4 convolutional blocks, alternating 2D-upsampling layers and 2D convolutional layers, **decreasing** number of filters.
- Activations: LeakyReLU (no activation at the output layer)
- Loss functions:
  - MSE loss at the **output** of the network, between the true and estimated spectrograms.
  - KL loss at the **distribution** of the encoded embeddings, between the learned distribution and the spherical isotropic Gaussian, $N(0, I)$.

Net. Block	N_{filt}	(k_x, k_y)	(p_x, p_y)	(u_x, u_y)
Conv_DS1	N	(5,5)	(2,2)	-
Conv_DS2	2N	(5,5)	(4,2)	-
Conv_DS3	4N	(5,5)	(4,4)	-
Conv_DS4(μ)	8N	(4,4)	(4,4)	-
Conv_DS4(σ)	8N	(4,4)	(4,4)	-
Sample	-	-	-	-
Conv_US1	4N	(5,5)	-	(4,4)
Conv_US2	2N	(5,5)	-	(4,4)
Conv_US3	N	(5,5)	-	(4,2)
Conv_US4	1	(5,5)	-	(2,2)

30th EUSIPCO EUROPEAN SIGNAL PROCESSING CONFERENCE
BELGRADE/SERBIA
29 AUG - 2 SEP 2022

Experimental Setup

- Experiments for both **personalized** (separate model for each patient) and **universal** (one model for all patients) cases.
- Baseline: The CAE model presented in [3].
- N = 32 filters at the outer convolutional layer, for both CAE and CVAE models.

Training Details:

- 5-fold cross-validation, data from the same session are assigned to the same fold.
  - Training: Only data from time periods where the patient condition was **stable**
  - Testing: Mixture of data from **stable** and **anomalous** (pre-relapsing or relapsing) time periods.
- Adam (lr=0.0003), batch size of 8.
- 200 epochs maximum, early stopping applied at 10 epochs.
- Loss weights: $W_{MSE} = 1$, and $W_{KL} = 0.01$

- Evaluation Metric: Mean **ROC-AUC score** over all sessions.

[3] C. Garoufis, A. Zlatintsi, P. Filntisis, N. Efthymiou, E. Kalisperakis, T. Karantinos, V. Garyfalli, L. Mantonakis, N. Smyrnis, and P. Maragos, "An Unsupervised Learning Approach for Detecting Relapses from Spontaneous Speech in Patients with Psychosis," in Proc. BHI 2021

Experimental Setup – Ablation Studies

- Probe point for the anomaly scores:
  - CAE: Output reconstruction MSE
  - CVAE: Output reconstruction MSE and input embedding KL divergence.

- Temporal aggregation function of the per-session anomaly scores:
  - Average pooling (AP): $S = \frac{1}{N}\sum_{i=1}^{N} s_i$
  - Max pooling (MP): $S = \max(s_i)$
  - Norm pooling (NP): $S = \frac{1}{N}\left(\sum_{i=1}^{N} |s_i|^p\right)^{1/p}$, p = 10.

- Normalization scheme (universal models):
  - Global normalization, i.e. a **shared** normalization transform for all patients.
  - Per-patient normalization, i.e a **separate** normalization transform for each patient.

30th EUSIPCO EUROPEAN SIGNAL PROCESSING CONFERENCE
BELGRADE/SERBIA
29 AUG – 2 SEP 2022

Results: Personalized Models

Average per-patient ROC-AUC score for the discrimination between sessions that correspond to stable, or anomalous, condition, for both CVAE and CAE personalized models.

Pooling Function	CAE [24]	CVAE	
		MSE	KL
AP	**0.668** ± 0.035	**0.673** ± 0.055	0.653 ± 0.052
MP	0.608 ± 0.060	0.617 ± 0.051	0.659 ± 0.045
NP	0.627 ± 0.058	0.640 ± 0.049	**0.678** ± 0.049

- The performance of the proposed CVAE is **comparable** to that of the deterministic CAE in the personalized case.
- No statistically significant difference ($p > 0.05$) between models.
- Average pooling performs the best when using the reconstruction MSE as the anomaly score for both CAE and CVAE models.
- Norm pooling gives the best results when using the KL divergence as the anomaly score.

Per-patient ROC-AUC scores for the discrimination between sessions that correspond to stable, or anomalous, condition, for both CVAE and CAE personalized models.

Patient ID	CAE [24]	CVAE	
		MSE	KL
#1	0.546 ± 0.069	**0.547** ± 0.081	0.523 ± 0.134
#2	**0.448** ± 0.093	0.418 ± 0.182	0.388 ± 0.120
#3	**0.711** ± 0.119	0.700 ± 0.159	0.656 ± 0.187
#4	0.676 ± 0.066	0.660 ± 0.034	**0.768** ± 0.066
#5	0.781 ± 0.053	**0.800** ± 0.095	0.774 ± 0.040
#6	0.512 ± 0.067	0.520 ± 0.173	**0.696** ± 0.083
#7	0.877 ± 0.076	**0.940** ± 0.074	0.720 ± 0.186
#8	0.800 ± 0.187	0.850 ± 0.292	**0.900** ± 0.200
Average	0.668 ± 0.035	0.673 ± 0.055	**0.678** ± 0.049

- Per patient results: ROC-AUC score **above 0.75** for 4/8 patients, **above 0.65** for 6/8

Results: Universal Models

- The CVAE model **outperforms** by a large margin the baseline CAE, especially when obtaining the anomaly score from the KL divergence.
- Application of **per-subject normalization** leads to performance equivalent to the one achieved by the personalized models.
- Statistically significant improvement ($p < 0.05$) over the baseline when using personalized normalization and the KL divergence as anomaly score.
- **Norm pooling** appears to perform the best as a temporal aggregation function.
- CVAEs perform better at a patient-independent setting -> speaker-invariant representations? [6]

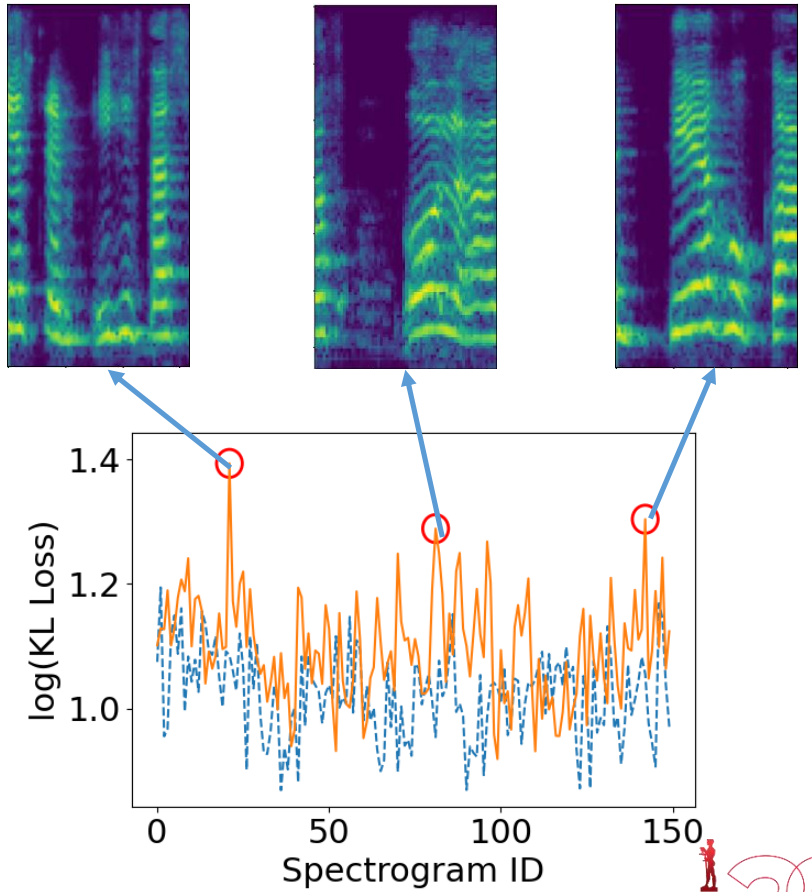
Average ROC-AUC score for the discrimination between sessions that correspond to stable, or anomalous, condition, for both CVAE and CAE universal models.

Pers. Norm	Pool. Func.	CAE [24]	CVAE	
			MSE	KL
✗	AP	0.504 ± 0.032	0.502 ± 0.016	**0.532** ± 0.023
✗	MP	0.542 ± 0.024	0.551 ± 0.023	**0.592** ± 0.031
✗	NP	0.531 ± 0.034	0.527 ± 0.024	**0.581** ± 0.036
✓	AP	0.552 ± 0.036	0.585 ± 0.021	**0.646** ± 0.035
✓	MP	0.541 ± 0.027	0.622 ± 0.034	**0.685** ± 0.040
✓	NP	0.542 ± 0.028	0.618 ± 0.030	**0.698** ± 0.042

[6] J. Chorowski, R. Weiss, S. Bengio, and A. van den Oord,"Unsupervised Speech Representation Learning using WaveNet Autoencoders," IEEE/ACM Trans. on Audio, Speech, and Language Process., vol. 27, no. 12, pp. 2041–2053, 2019.

Results: Qualitative Analysis

- KL scores for two interview sessions over time:
  - Stable condition (dashed blue)
  - Relapsing condition (orange).



Observations:

- Not significantly higher KL scores during the relapsing session
- Appearance however of **a few peaks** (in red circles)
- Aural inspection of the respective segments -> **abrupt in-utterance disruptions** of the patient's speech flow.

30th EUSIPCO EUROPEAN SIGNAL PROCESSING CONFERENCE
BELGRADE/SERBIA
29 AUG – 2 SEP 2022

Conclusions & Future Work

Explored the potential of Convolutional Variational Autoencoders (CVAEs) in speech-based relapse detection prediction in psychotic patients.

- Personalized case: **Comparable** performance to a CAE baseline.
- Universal case: **Significant improvement** over CAEs, in conjunction with a **personalized normalization** scheme.

What's next?

- Utilization of **multimodal information**, such as text transcripts, or data collected from smartwatches.
- Taking advantage of **longer-term dependencies** in interviews:
  - During the same utterance.
  - In successive utterances.

30th EUSIPCO EUROPEAN SIGNAL PROCESSING CONFERENCE
BELGRADE/SERBIA
29 AUG - 2 SEP 2022

Thank you for your attention!

This research has been financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE (project acronym: e-Prevention, code:T1EDK-02890/ MIS: 5032797)
For more information and project results, visit https://eprevention.gr

e-Prevention