The information-differentiated loss function for speech feature clustering in a low-resource environment

Viacheslav Kovtun[1,†], Victoria Vysotska[2,*,†] and Oksana Kovtun[3,†]

[1] *Vinnytsia National Technical University, Khmelnytske shose, 95, Vinnytsia, 21021, Ukraine*
[2] *Lviv Polytechnic National University, Stepan Bandera Street, 12, Lviv, 79013, Ukraine*
[3] *Vasyl' Stus Donetsk National University, 600-richchya Str., 21, Vinnytsia, 21000, Ukraine*

Abstract

The article investigates the problem of forming a portable cluster structure in the latent space of language representations without using annotated data, which is especially relevant for zero-shot classification tasks, low-resource language processing, and generalization to new domains. Based on a critical review of modern approaches to unsupervised cluster learning, a loss function is proposed that combines global entropy regularization with scaling of the contribution of examples depending on the level of model confidence. The value of the scaling parameter is determined automatically based on the local decrease in the entropy of the cluster distribution, which serves as an indicator of the isolation of a language segment in the latent space. Such a mechanism makes it possible to suppress the contribution of latently unstable examples without removing them, ensuring structural adaptation of the cluster topology in new domains without retraining. Experimental results on GlobalPhone, CommonVoice and unseen-domain Ukrainian Speech Corpus demonstrated a reduction in average cluster entropy to 0.88, suppression of over 60% of unstable segments and an increase in cluster structure consistency by 19% in zero-shot mode. The proposed approach provides stable and adaptive clustering in the absence of annotations, in particular in cold start scenarios and rapid structuring of speech data in new environments.

Keywords

unsupervised cluster learning, entropy regularization, loss scaling, structural adaptation, latent space, language representations, cluster portability, zero-shot inference

1. Introduction

One of the current problems of modern computational linguistics is the formation of stable and structured clusters in the latent space of language representations, in particular, such as acoustic prototypes or pseudo-phonemes, which is critically important for the tasks of automatic speech analysis in the absence of labels. This task underlies a wide range of scenarios, including zero-shot classification at the level of speech segments [1, 2], the transfer of acoustic models between languages or domains with significant differences in style, diction or acoustic conditions [3], the construction of pseudo-phonemic inventories [4] and the structuring of speech corpora for low-resource languages [5, 6]. Such tasks arise, for example, in the construction of speech recognition systems for Arabic dialects, the transliteration of names in multilingual chatbots, the clustering of sounds in audio data from field studies of indigenous languages or the filtering of noisy speech fragments in media content. In the context of Ukraine, this also has applied significance: in particular, in the creation of automated speech processing systems for Western Ukrainian dialects, processing audio data from social networks and public speeches in wartime conditions, where the recording quality is unstable, or in the formation of a basic cluster structure for building language support for the Crimean Tatar and Gagauz languages, which are underrepresented in the digital environment.

ISW-2025: Intelligent Systems Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine
* Corresponding author.
† These authors contributed equally.
vkovtun@iitis.pl (V. Kovtun); victoria.a.vysotska@lpnu.ua (V. Vysotska); o.kovtun@donnu.edu.ua (O. Kovtun)
0000-0002-7624-7072 (V. Kovtun); 0000-0001-6417-3689 (V. Vysotska); 0000-0002-9139-8987 (O. Kovtun)

In such contexts, there is a need for models that can form an ordered latent structure without external labels and with minimal intervention. Although self-supervised learning and contrastive methods have given significant impetus to the development of representations in computational linguistics, most of them are either based on global heuristics or do not take into account local variability in the confidence of the model. In particular, common approaches do not include mechanisms that would allow taking into account the instability of individual examples when calculating losses and are not focused on ensuring the portability of the cluster structure in new domains. It remains insufficiently studied how global entropy regulation can be combined with local adaptation without external support, especially in conditions of substantial domain change. The need for such approaches is especially acute in the field of automatic processing of low-resource languages, where adaptation to new language environments must occur without prior retraining and without relying on linguistic annotations. The lack of established methods that combine entropic compaction of cluster structure with dynamic latent control at the level of individual segments creates a significant scientific gap. This gap determines the feasibility of research aimed at the systematic study of loss functions capable of implementing self-regulating adaptation in clustering problems of computational linguistics.

One of the key challenges in the clustering of unlabelled language representations is the formation of a latent space that is simultaneously structured, stable, and transferable to new domains. In response to this problem, several research directions have emerged in computational linguistics and machine learning, involving different approaches to organizing the internal space of a model without external control: contrastive learning, entropy regularization, pseudo-labelling, prototype-based learning, curriculum learning, and confidence-based loss scaling. Each of these approaches has its own motivation, application mechanisms, and certain limitations, overcoming which is the basis for further research. The following is a critical review of the main of these strategies, taking into account their potential and vulnerabilities, which directly determine the need for new solutions.

Contrastive learning [7, 8] involves organizing the latent space by training on positive and negative pairs of representations. The model learns to reduce the distance between examples that are considered similar (for example, augmented versions of the same speech fragment) while increasing the distance between examples from different sources. This approach is the basis of many modern self-supervised systems, in particular SimCLR [9], Wav2Vec 2.0 [10] and HuBERT [11], which have shown high efficiency in feature detection tasks without annotations. In the context of the problem of constructing a portable cluster structure without labels, contrastive learning allows you to create a well-organized representation space that distinguishes categories. Its application in speech processing enables you to cluster segments according to their acoustic similarity without being tied to phonemic labels. The main advantage of this approach is the ability to form discriminative representations without redundant hyperparameters or the need for annotation. However, contrastive learning usually operates with the global structure of the space and does not take into account local latent uncertainty. All examples lose or gain the same contribution to the loss function, regardless of their stability or position with respect to the prototypes. In the case of noisy or unpredictable data, this can lead to a violation of cluster integrity since the model does not have a built-in mechanism for suppressing dubious examples. This drawback is one of the key entry points for justifying the need for adaptive loss scaling, implemented in our study through a mechanism $\alpha(\vec{r})$ that allows local reduction of the influence of latently unstable fragments without rigidly excluding them.

Entropy regularization [12, 13] involves adding a special term to the loss function, which is aimed at reducing the entropy in the model's output distributions. The main idea is that the model should strive for "decisive" (low-entropy) predictions, even in the absence of labels. One of the most famous classical approaches is entropy minimization, where entropy minimization is used to strengthen the determinism of the classifier in semi-controlled conditions. In speech processing, similar techniques are used, for example, to order the discretization spaces in models such as HuBERT or APC (autoregressive predictive coding) [14]. In the context of the problem of constructing a portable cluster structure, this approach allows for reducing excessive uncertainty in relation to prototypes

or clusters, forming a clearer internal structure in the model. In practice, this means that examples with lower entropy will have an advantage in forming centroids or strengthening the boundaries between clusters. The main advantage of this approach is its universality and simplicity: the entropy term is easily integrated into most loss functions, and its minimization often contributes to improving cluster integrity. At the same time, this approach has a significant limitation - it acts equally on all examples, regardless of whether their uncertainty is a consequence of noise, mixing or latent uninformativeness. The lack of local control means that the model can artificially reduce entropy, even for examples that do not have an explicit cluster nature, thereby distorting the structure of the space.

Pseudolabeling [15, 16] is a strategy in which the model itself generates labels for the raw data based on its predictions. These labels are then considered "conditionally correct", and the model is trained on them, usually in a supervised manner. This approach is widely used in semi-supervised learning, for example, in FixMatch [17] or Noisy Student [18] methods. In the field of speech processing, pseudolabeling is used, in particular, in the clustering phases of models such as HuBERT or TERA, where the first passes of clustering generate "soft targets", which are then used as the basis for training subsequent layers. In the task of building a portable cluster structure, pseudo-labeling allows the model to refine the classification of input examples step by step, focusing on the most confident predictions. It will enable it to gradually "correct" the unstructuredness of the latent space and bring it closer to a more ordered form. Among the advantages of this approach are its flexibility, ability to accumulate knowledge, and support for gradual self-organization even in the absence of external annotations. However, it also has significant drawbacks. First, pseudo-labeling usually requires a hard confidence threshold, below which examples are not used, which means a large amount of data is lost. Second, this approach lacks a built-in mechanism for soft control over the contribution - an example is either accepted for training in full or rejected. Such a binary nature makes the model sensitive to errors in the early stages and does not allow for flexible suppression of the influence of dubious segments.

Prototype learning [19] involves organizing the latent space around a fixed or dynamically updated set of centres – the so-called prototypes, which act as representative points for clusters or classes. In this approach, each example in the projection space approaches one of these prototypes, and the loss function itself usually optimizes the distance to the nearest centre. Modern implementations include [20] DeepCluster, SwAV, DINO, and in speech processing – HuBERT, DeCoAR [21] and WavLM [22], where clusters obtained via k-means or GMM serve as internal labels for subsequent training iterations. In the context of building a transferable cluster structure, these approaches have a significant advantage: they encourage the model to form compact regions around the centres, which is well consistent with the intuition of the latent space as a set of semantically similar units. In addition, the cluster structure becomes more interpretable and easily portable since prototypes can be used as a basis for classification in a zero-shot mode. However, even with these advantages, classical prototype learning has its limitations. First, all examples participate in the formation of losses with the same weight, which means that it is impossible to ignore or suppress latently unreliable examples. Second, the updating of prototypes is usually performed without taking into account the confidence of belonging to the cluster, which can lead to the displacement of the centres under the influence of noisy or poorly classified points. Many implementations also lack the means to isolate cases with latent ambiguity and, therefore, do not provide adaptive selectivity.

Curriculum learning [23, 24] or self-paced learning [25] involves the model initially focusing on "simple" examples, gradually moving to more complex ones. The idea is that an orderly input of information – from easy to difficult – contributes to better generalization and stability of the model. This principle has been applied in various contexts, from computer vision to NLP [26, 27], as well as in clustering, where complexity can be determined, for example, by the distance to the prototype, the entropy of the output distribution or the stability of the prediction. In language tasks, such approaches have been used in adaptive variations of APC or in training multilingual acoustic models. In the context of building a portable cluster structure, these approaches have a strong intuitive motivation: they allow the model to form the backbone of the cluster topology on "reliable" examples

before encountering latently ambiguous zones. Thanks to control over the order of inclusion of examples in training, it is possible to avoid premature re-adaptation or deformation of the space. However, the implementation of curriculum/self-paced learning has significant practical drawbacks. The most important one is the need for an external assessment of the "complexity" of the example, which often involves the use of an auxiliary module, manual sorting, or additional hyperparameters. In addition, complexity in a multidimensional latent space is a fuzzy and dynamic concept: an example that is complex at the first stage may become simple later, and vice versa. Ше makes it challenging to integrate such approaches into unlabeled clustered learning.

As the above review shows, none of the existing approaches to building an unsupervised cluster structure provides a full-fledged combination of global structuring of the latent space with local adaptive selectivity at the level of individual examples. Contrastive learning and prototype methods allow for forming an ordered space but do not take into account uncertainty in predictions and are unable to reduce the impact of latently unstable data. Entropy regularization provides global control over fuzziness but operates without taking into account the context or nature of a specific example. Pseudolabeling and self-paced learning provide selectivity but require hard thresholds or external intervention, which limits their flexibility and portability. At the same time, methods that scale losses based on model confidence remain underexplored, especially in the field of speech processing. This combined drawback - the lack of an internally adaptive, continuous mechanism for controlling the contribution of examples to the loss function while maintaining cluster integrity - determines the relevance of this study. Exploring the possibility of simultaneously implementing entropy compaction and local latent scaling responds to the scientific community's request for creating loss functions capable of maintaining robustness, selectivity, and portability in a clustered, label-free space.

The object of the study is the process of forming an adaptive cluster structure in the latent space of language representations in the absence of explicit annotations, taking into account the local confidence of the model and the need to ensure the portability of this structure to new domains without additional training.

The subject of the study is a set of approaches to constructing loss functions for unsupervised learning, in particular, the methods of entropy regularization, prototype representation, pseudolabeling and latent-guided scaling, which ensure the adaptive formation of a cluster structure and its portability to new domains without the use of explicit annotations.

The purpose of the study is the theoretical justification and experimental verification of the loss function for unsupervised cluster learning, which combines global entropy regularization and local scaling of losses based on latent confidence in order to ensure the structuredness, selectivity and portability of the cluster organization in the latent space of language representations.

2. Models and methods

2.1. Research Statement

In the context of constructing a differentiated information loss function for neural network speech models under resource-constrained conditions, a formalized approach to processing variable acoustic realizations of phonemes is key. Given that the articulatory and acoustic realization of each phoneme is a stochastic function, dependent on both the individual characteristics of the speaker and the noise context, the cluster model of phonemes acquires fundamental importance.

Let $C \in \mathbb{N}$ denote the total number of phoneme classes (clusters) that the system must identify or train. Each cluster $c \in \{1,2, \dots, \mathrm{C}\}$ corresponds to a set of vectors $R_c = \{\vec{r}_{ci} \in \mathbb{R}^n | i = 1, \dots, I_c\}$, where $I_c \in \mathbb{N}$ is the number of available realizations of the c-th phoneme and $\vec{r}_{ci}$ is the feature vector for the i-th realization of this phoneme. Each vector $\vec{r}_{ci}$ represents an elementary speech unit obtained after preliminary speech processing (for example, spectral or LPC decoding).

Each cluster R_c is embedded in a latent space on which the information centre (prototype) $\vec{r}_c^*$ is defined, which minimizes the generalized Kullback–Leibler divergence with respect to all elements

of the cluster: $\vec{r}_c^* = arg\max_{\vec{z}\in R_c} \sum_{i=1}^{I_c} D_{KL}\left(P_{\vec{r}_{ci}} \| P_{\vec{z}}\right)$, where $P_{\vec{r}}$ denotes the estimate of the probability distribution for the feature vector $\vec{r}$, which can be empirical or parametric (for example, a normal distribution with a covariance matrix estimated from a sample). Such minimization is implemented in the form of a loss function of the KL-Loss type in the process of optimizing the neural network.

The input feature vector of the speech signal $\vec{r} \in \mathbb{R}^n$ obtained at the current processing step corresponds to one of the clusters R_v, where the index v is determined by the rule of least divergence: $v = arg\min_{c\in\{1,\dots,C\}} D_{KL}\left(P_{\vec{r}} \| P_{\vec{r}_c^*}\right)$. Interpreting this procedure as soft metric learning, we get the opportunity to form a latent space in which heuristically defined phoneme prototypes allow us to implement a stable and differential comparison of speech signals based on information proximity.

On the basis of the input, each phoneme is represented as an informationally consistent cluster of vectors R_c, and the problem of speech quality analysis is transformed into the issue of optimal classification distribution of elements of the space $\mathbb{R}^n$ by C statistically justified centers $\{\vec{r}_c^*\}_{c=1}^C$, with subsequent training in the statistical classification of signals by a teacher.

To ensure a stable classification of phonemic representations in a differentiated speech processing model under resource-constrained conditions, it is advisable to formulate a criterion for assigning the current segment to one of the $C \in \mathbb{N}$ phonemic classes in terms of information discrepancy. It is assumed that the distribution of feature vectors $\vec{r}_{ci} \in \mathbb{R}^n$, which form the cluster $R_c = \{\vec{r}_{ci}\}_{i=1}^{I_c}$, is approximated by a multivariate normal law with zero mean and covariance matrix $\mathrm{K}_{ci} \in \mathbb{R}^{n\times n}\ddot{x}$. This approach allows us to describe the stochastic nature of speech, which is key in resource-constrained conditions.

The current input vector $\vec{r} \in \mathbb{R}^n$, evaluated within the local window, belongs to one of the clusters R_c according to the generalized Kullback–Leibler divergence criterion between the empirical distribution $\hat{P}_{\vec{r}} \sim \mathrm{N}\left(0, \widehat{\mathrm{K}}\right)$ and the prototype distribution $P_{\vec{r}_{ci}} \sim \mathrm{N}(0, \mathrm{K}_{ci})$. The expression gives the corresponding divergence:

$$\theta_c(\vec{r}) = \frac{1}{2n}\left[\mathrm{tr}\left(\frac{\widehat{\mathrm{K}}}{\mathrm{K}_{\vec{r}}}\right) - logdet\left(\frac{\widehat{\mathrm{K}}}{\mathrm{K}_{\vec{r}}}\right) - n\right], \quad (1)$$

where $\widehat{\mathrm{K}}$ is the estimate of the autocovariance matrix for $\vec{r}$, calculated using a fixed-length sliding window, and $\mathrm{tr}(\cdot)$, $det(\cdot)$ are the trace and determinant, respectively. In this definition, expression (1) is fully differentiable and is easily embedded in gradient deep learning algorithms.

Next, for each cluster R_c, a matrix of pairwise information discrepancies is formed:

$$\Theta_{ij}^{(c)} = \theta\left(\vec{r}_{ci}, \mathrm{K}_{cj}\right) = \frac{1}{2n}\left[\mathrm{tr}\left(\frac{\widehat{\mathrm{K}}_{ci}}{\mathrm{K}_{cj}}\right) - logdet\left(\frac{\widehat{\mathrm{K}}_{ci}}{\mathrm{K}_{cj}}\right) - n\right], i \geq 1, j \leq I_c, \quad (2)$$

where $\widehat{\mathrm{K}}_{ci}$ is the empirical autocovariance matrix for the vector $\vec{r}_{ci}$, and K_{cj} is the covariance matrix of the corresponding realization $\vec{r}_{cj}$. The matrix $\Theta^{(c)} \in \mathbb{R}^{I_c\times I_c}$ serves as the basis for determining the information centre of the corresponding cluster.

The definition of the information prototype $\vec{r}_c^* \in R_c$ is based on the minimization of the total divergence

$$\vec{r}_c^* = \vec{r}_{c\kappa}, \kappa = arg\min_{j\in\{1,\dots,I_c\}} \sum_{i=1}^{I_c} \Theta_{ij}^{(c)}. \quad (3)$$

That is, the implementation $\vec{r}_{c\kappa}$ is chosen as the information centre, which is, on average, the closest to all other implementations in terms of the generalized divergence (2). It allows us to specify a phoneme representative that maintains the highest consistency with different implementations of this phoneme.

Formulas (1)–(3) form a coherent information-theoretic basis for training a neural network model in the formulation of metric-based classification of speech signals. In combination with a training supervisor, these expressions allow us to perform preliminary training on labelled data with phonemic annotation, which can then be adapted to new conditions or speakers.

2.2. Formalization of the information loss function for neural network clustering of speech features under resource-constrained conditions

In the tasks of automatic speech modelling in a low-resource environment, it is critically important to ensure the ability of the system to independently detect the latent structure of phoneme-like units without prior annotation. In such cases, the neural network model should implement adaptive clustering of feature vectors with the possibility of gradually increasing the number of clusters as new data arrives. This approach is interpreted as a stochastic sequence of conditionally supervised tasks with a variable number of clusters - through the introduction of an information $(C+1)$-element as a structure for recursive updating of the classification space.

Let the speech signal be represented as a discretized series of amplitudes $R(t) = \{r_1, r_2, \dots, r_L\} \subset \mathbb{R}$, segmented with a fixed time step $\tau \in [5{,}15]$ ms. Each segment of length $L \in \mathbb{N}$ is converted into a feature vector $\vec{r}$ using, for example, LPC or MFCC decoding.

Let us denote the first segment as $\vec{r}_1$ and initialize the cluster $R_1 = \{\vec{r}_1\}$ with the covariance matrix $\mathrm{K}_1^{(1)} = \widehat{\mathrm{K}}_1$, estimated as $\widehat{\mathrm{K}}_1 = \frac{1}{L-1}\sum_{t=1}^{L}(\vec{r}_1(t) - \mathrm{E}[\vec{r}_1])(\vec{r}_1(t) - \mathrm{E}[\vec{r}_1])^T$, where $\mathrm{E}[\vec{r}_1]$ is the mean value of the vector $\vec{r}_1$. The first cluster forms the prototype set, and the number of clusters is set as $C = 1$.

Let $\vec{r}_2 \in \mathbb{R}^n$ be the next segment shifted in time by τ. We calculate the generalized Kullback–Leibler divergence (or its parametric variant) between its distribution and the first cluster:

$$\theta(\vec{r}_2, R_1) = \frac{1}{2n}\left[\mathrm{tr}\left(\frac{\widehat{\mathrm{K}}_2}{\mathrm{K}_1}\right) - logdet\left(\frac{\widehat{\mathrm{K}}_2}{\mathrm{K}_1}\right) - n\right], \tag{4}$$

where K_2 is the covariance matrix for $\vec{r}_2$, which can be estimated as a learnable head or batch-normalized calculation with a fixed window order. In practical scenarios, in particular, in high noise conditions, it is permissible to generalize (4) to α-divergences or Sinkhorn divergences to improve stability.

The vector $\vec{r}_2$ is included in R_1 if the inequality holds

$$\theta(\vec{r}_2, R_1) \leq \theta_0, \tag{5}$$

where $\theta_0 \in \mathbb{R}_+$ is a learnable threshold that can be implemented as a parameter that is optimized during training through sigmoidal relaxation. If (5) is not satisfied, a new cluster $R_2 = \{\vec{r}_2\}$ is created, and $C \leftarrow C + 1$.

To ensure the reliability of clusters, regularization is introduced using the threshold $L_0 \in \mathbb{N}$, which sets the minimum allowable total duration of the cluster:

$$|R_c|\tau \geq L_0. \tag{6}$$

Clusters that do not satisfy (6) are considered marginal and do not participate in the further construction of the phonetic database. Such heuristics are critical in low-resource conditions, for example, when creating an offline speech access system in the Ukrainian language in field or military conditions, where the reliability of clustering is critical.

The adaptive procedure (4)–(6) provides online clustering of the speech stream with a dynamic number of clusters $C^* \leq C$ forming a structured set $\{R_c\}_{c=1}^{C^*}$. This set represents the speaker's latent phoneme space. It serves as a basic prototype layer for the subsequent differentiated loss function - both in the form of information-theoretic and contrastive loss, oriented to preserving the separation of phonemes in the embedding spaces.

In the developed system of online clustering of speech segments, it is crucial not only to identify structurally stable clusters but also to provide a normalized spectral representation that minimizes the influence of individual acoustic variations of the speaker. For this purpose, a modified version of autoregressive spectral normalization is used, adapted for the needs of differentiated learning in a neural network environment. The key element in this is the processing of each speech segment $\vec{r}(t) \in \mathbb{R}^n$ as an implementation of a low-order stationary process, which is approximated by the

autoregressive (AR) model $\vec{r}(t) = \sum_{k=1}^{K} a_k \vec{r}(t-k) + \vec{\varepsilon}(t)$, $t \in \mathbb{Z}_+$, where $K \in \mathbb{N}$ is the order of the model; $\{a_k\}_{k=1}^{K} \subset \mathbb{R}$ is a set of learnable AP coefficients; $\vec{\varepsilon}(t)$ is a residual noise vector. The constructed vector $\vec{a} = [a_1, \dots, a_K]^T$ serves as a latent feature representation of the speech segment and is invariant to the absolute energy and timbre structure of the signal. It can be used as an input to the clustering head of the model or as a component in self-supervised pretext tasks (e.g., temporal ordering or frame prediction).

To estimate the spectral difference between the vector $\vec{a}^{(r)}$ of the current segment and the prototype $\vec{a}^{(c)}$ of cluster c, a normalized weight function is introduced:

$$\theta_c\left(\vec{a}^{(r)}\right) = \sum_{f=1}^{F} \left[\frac{1}{1+\pi(f)} \left(\frac{1}{K} \sum_{k=1}^{K} \left(a_k^{(r)} - a_k^{(c)} \right)^2 \right) \right], \tag{7}$$

where $f \in \{1, \dots, F\}$ are frequency indices, and $\pi(f)$ is a spectral mask that determines the weight of each frequency channel. The mask $\pi(f)$ can be implemented as a fixed (Mel- or Bark-filtering) or as a learnable function adaptive to the speech domain, which allows the model to focus on perceptually relevant frequencies (for example, in the region of the first two formants).

Note that in order to increase the sensitivity to atypical spectral deviations, expression (7) can be generalized in the form of an exponential divergence:

$$\theta_c\left(\vec{a}^{(r)}\right) = log\left(\frac{1}{K} \sum_{k=1}^{K} exp\left(\lambda \left| a_k^{(r)} - a_k^{(c)} \right| \right) \right), \tag{8}$$

where $\lambda \in \mathbb{R}$ is a sensitivity parameter that enhances the effect of large spectral deviations, the divergence (8) provides increased resolution with limited data and is used as an internal module in the context of margin-based loss functions. For example, it can be integrated into the contrastive loss (CL) as

$$L_{CL} = y\theta_c + (1-y)\, max(0, \mu - \theta_c), \tag{9}$$

where $y \in \{0,1\}$ is a binary cluster-correspondence feature, and $k \in \mathbb{R}_+$ is a hyperparameter margin that sets a threshold for distinguishing clusters in the latent space.

Autoregressive spectral normalization provides effective invariance to loudness, timbre, and speaker voice parameters. It is achieved by smoothing out energy fluctuations induced by anatomical features of the speech tract, thereby preserving relevant phoneme dynamics. In resource-constrained scenarios, such as when developing ASR systems for the Ukrainian language based on field recordings, such invariance is critical to ensuring the generalizability of the model. Therefore, formulas (7)–(9) define a spectrally normalized divergence function that is fully differentiable, interpretable, and compatible with both supervised and self-supervised downstream-loss architectures. Vectors $\vec{a}^{(c)}$ formed on the basis of these divergences form the basis for constructing an information-optimal phonetic database.

In constructing a differentiated loss function for a neural network model of speech feature processing, an important step is to determine secondary clustering quality criteria that can serve as optimization meta-functions, regularizers, or latent structure coherence indicators. Of particular value in this context are information-theoretic metrics that describe the entropic organization of the set of clusters $\{R_c\}_{c=1}^{C^*}$ obtained as a result of recursive clustering (see expressions (4)–(9)).

The basic criterion is the empirical distribution of clusters:

$$P_c \frac{|R_c|}{\sum_{s=1}^{C^*} |R_s|}, \; c = \{1, \dots, C^*\}, \tag{10}$$

where $|R_c|$ is the number of speech segments that were assigned to cluster c based on minimizing the selected divergences $\{(4), (7), (8)\}$. Thus, $P = [P_1, \dots, P_{C^*}]^T \in \Delta^{C^*-1}$ is a probability vector that is a dynamic variable in the classification architecture and directly depends on the parameters of the coding space. The Shannon entropy over the distribution (10) defines a generalized cluster differentiation metric:

$$H(P) = -\sum_{c=1}^{C^*} P_c \log P_c, \tag{11}$$

which reaches a maximum of $H \log C^*_{max}$ in the case of a completely uniform distribution. This situation may indicate excessive differentiation - the model does not detect any dominant structure, which is especially undesirable in low-resource conditions, where phonemes are presented with different frequencies. In variant implementations, it is allowed to replace (11) with parametric variants, in particular, the Rényi or Tsallis entropy, which allows controlling the sensitivity to rare clusters.

To provide structural control over clustering, a normalized redundancy metric is introduced:

$$\Omega = 1 - \frac{H(P)}{\log C^*}, \tag{12}$$

which varies in the interval [0,1] and is interpreted as a structural concentration coefficient. Low values of $\Omega \approx 0$ indicate uniformity of the cluster space, indicating noisy or unstructured behaviour, while high values of $\Omega \to 1$ indicate a tendency to collapse into one or two dominant clusters. Thus, Ω balances between variability and over-aggregation. In practice, Ω is integrated into the primary loss function as a global entropy regularizer, regulating the complexity of the cluster distribution:

$$L_H = L_{CL} + \beta\Omega, \tag{13}$$

where $\beta \in \mathbb{R}_+$ is a hyperparameter that determines the regularization weight. The construction (13) allows for the avoidance of the collapse of the coding space and the maintenance of the dissimilarity of features without renormalization at the level of each pair. The regularizer Ω controls both excessive uniformity and imbalance in favour of frequent clusters, which is especially critical for low-resource languages, such as Ukrainian, where available corpora are usually phonologically unbalanced.

In general, formulas (10)–(13) form a single entropy-normalized criterion of clustering quality, which allows not only to evaluation the result of structuring the latent space but also to actively manage it within the gradient-oriented loss function.

In the framework of constructing a full-fledged information loss function for a neural network model of language feature processing, it is necessary to take into account not only the global characteristics of the cluster space but also local confidence indicators for each input segment. For this purpose, a specialized function is introduced that allows us to estimate the relative certainty of the classifier's decision based on the contrast between the divergence to the nearest centre and the average divergence to the remaining clusters. Formally, if $\theta_c(\vec{r})$ is the information or spectral divergence (defined according to $\{(4), (7), (8)\}$), and $= arg \min_c \theta_c(\vec{r})$ is the cluster to which the input segment $\vec{r} \in \mathbb{R}^n$ is assigned, then the local classification confidence metric is defined as:

$$\Gamma(\vec{r}) = \frac{1}{C^*-1} \sum_{\substack{c=1, \\ c \neq v}}^{C^*} \frac{\theta_c(\vec{r})}{\theta_v(\vec{r})}. \tag{14}$$

Expression (14) allows us to quantitatively assess the degree of dissimilarity: the lower the value of (14), the clearer the cluster boundary is observed for a given example. In practice, high values of this quantity indicate blurring, transient nature or the presence of noise artefacts that complicate classification. To take this information into account in a balanced manner during training, a sigmoidal function of the confidence scale is introduced, which forms the weight coefficient of the example based on the obtained level of classification uncertainty. Such a coefficient is given in the form:

$$\alpha(\vec{r}) = \frac{1}{(1+exp(\gamma(\Gamma(\vec{r})-\varphi)))}, \gamma, \varphi \in \mathbb{R}_+. \tag{15}$$

The scalar multiplier (15) acts as a local regulator of the influence of the example on the overall loss function: at high confidence (15), it approaches unity, and at fuzzy clustering, it approaches zero. The parameter φ determines the critical level of uncertainty that separates "useful" examples from potentially noisy ones, and γ regulates the steepness of the transition between confidence zones.

Based on these local dynamics, a final expression for the loss function is formed, which takes into account both the local weight of the example and the global entropy regularizer of the cluster space. The final function takes the form:

$$L_{fin} = \alpha(\vec{r})L_{CL} + \beta\Omega. \quad (16)$$

It is important to emphasize that, unlike function (13), which models global structural balance, function (16) is locally adaptive and provides flexible modulation of the influence of each specific example based on its position in the latent space. This approach allows the model to effectively suppress the influence of segments that are uncertain or marginal while maintaining the informativeness of clustering in the central zones. It is vital in resource-limited conditions, where processing may be accompanied by a high proportion of unpredictable acoustic variations, and the structure of the phoneme space may be incomplete or domain-dependent. The resulting loss function (16) provides adaptability, noise resistance, and structural coherence without the need for a rigid supervisor or a complete phoneme corpus.

3. Results and Discussion

In speech systems aimed at use in resource-constrained environments, in particular, in field or mobile applications of the Ukrainian language, traditional supervised methods quickly exhaust their potential. The high level of acoustic noise, the lack of annotated data, the unpredictability of speaker variations, and the domain instability of signals make it necessary to rethink the principles of clustering speech features. It is critical that the model not only processes the signal but also independently structures the latent feature space with an internally consistent organization while maintaining sensitivity to phoneme-like units without an external supervisor. In this context, this section is focused on empirically testing the effectiveness of the information-differentiated loss function, formalized in expression (16), as the architectural core of the clustering model for short speech segments. The uniqueness of this function lies in the combination of three complementary components: contrastive spectral divergence to ensure discriminability, entropy control Ω to balance the cluster space, and a local confidence scaling mechanism $\alpha(\vec{r})$, which dynamically reduces the influence of border or artefact segments. Together, they form an adaptive loss function that provides not only consistent clustering but also interpretability of the structure, noise resistance, and transferability between speech domains. The study hypothesizes that such a composition of functional modules allows for the formation of a coherent latent structure capable of generalization, even in cases of complete absence of phonemic marking. It is expected that the trained model will be able not only to effectively structure signals within the corpus on which it was trained but also to maintain topological stability when transferred to new speech domains without further training or recalibration of prototypes. The effectiveness of this approach is considered not as a formal increase in accuracy but as an opportunity to create a new type of speech system - those that work in real-time, without connection to external bases, with a critically low resource threshold. Such systems are relevant in the context of building Ukrainian-language voice access interfaces for military use, crisis response, or humanitarian support in areas with limited infrastructure.

In situations where speech resources are limited in both volume and quality, the model must be trained in an unstable, domain-inhomogeneous speech stream. For this study, two contrasting Ukrainian-language audio corpora were selected. Ukrainian GlobalPhone represents voiceover speech in a controlled acoustic environment and serves as a conditional standard against which the model's ability to organize clustering is tested. In contrast, CommonVoice (uk) captures everyday amateur recordings with their inherent irregular noise, background distortion, and voiceover variability—that is, it simulates a typical low-resource scenario typical of field applications.

Each signal is divided into overlapping windows of 25 ms duration with a step of 10 ms, which corresponds to the standard segmentation of speech at the microstructural level. For each window, a log-Mel spectrogram of dimension 64 is calculated. This type of feature was chosen because of its consistency with the requirements of the information-differentiated loss function: the spectrogram

allows us to introduce parametric estimates of spectral discrepancy (reversals (7)–(9)) into the contrast matching module. Before the features enter the clustering part of the architecture, they undergo spectral normalization, implemented as a low-order AR filter. This operation suppresses the influence of timbre, loudness, and other speaker-dependent characteristics, leaving only those frequencies that are relevant from the point of view of phonemic differentiation. It is at this level that the local confidence mechanism $\alpha(\vec{r})$ is introduced, which relies on normalized representations to assess the stability of the classification (expressions (14)–(15)). As a result, the model receives two streams of input data that differ in domain nature but are unified in terms of the processing procedure. It allows us to test not only the clustering itself but also the ability of the loss function (16) to preserve the structural logic within one corpus and transfer it to another - without supervision, retraining, or recalibration. Such a formulation allows us to evaluate the effectiveness of information adaptation not at the level of an artificial metric but as a fundamental cognitive strategy for structuring speech under constraints that correspond to real usage scenarios.

The model takes as input a sequence of normalized log-Mel spectrograms, which are fed to a two-layer BiLSTM encoder with 128 units. The use of a two-way recurrent structure is explained by the need to take into account the microcontext of the signal: the cluster affiliation of each segment is determined not in isolation but within the framework of a local dynamic template that reflects the natural nonlinearity of the phonetic stream. At the output of BiLSTM, a feature vector is formed, which passes through the learnable block of autoregressive spectral normalization. This block performs the function of smoothing acoustic variations caused by speaker features or distortions, which is especially important for field use scenarios. In the centre of the model is a learnable layer with cluster centres, the number of which is limited by the value $C = 20$. This decision is not dictated by convenience but is fundamental in the low-resource context: the model must learn to recognize a repeating structure without unlimited expansion of the cluster space. The centres are randomly initialized and adapted in the optimization process, which is guided by the loss function (16). Instead of classically calculating gradients based only on the error, here, the cumulative divergence (contrast, entropy and scale) between each example and the reference prototypes of the latent space is minimized. Unlike traditional models, where the loss function is an external means of error control, this architecture acts as an internal mechanism for structure formation. Each of its components is integrated at the parameter level: the contrastive divergence (7)–(9) is tied to learnable clusters, the entropy regularizer (10)–(13) models the distribution of examples, and the scaling factor (14)–(15) acts as an adaptive filtering of signals with low classification certainty. As a result, formula (16) becomes not just an optimization function but a rule for organizing the latent space that occurs inside the model itself.

The model is trained using the Adam optimizer at a rate of 10^{-3}, a batch size of 64, and a horizon of 100 epochs. None of the elements are supported by supervision – clustering occurs in the absence of any annotation. It is not a limitation but a strategy: in conditions where language resources are limited, an external supervisor or resuscitation of models through retraining may be impossible.

For qualitative diagnostics of the order of the latent representation space formed by the neural network during training, the t-SNE method was used to project multidimensional vectors into a two-dimensional space. This visualization allows interpretation of the topology of the cluster space: the degree of compactness, the presence of overlaps, and the distance between segment groups. The visualizations were performed separately for the Ukrainian GlobalPhone (Fig. 1) and CommonVoice (uk) (Fig. 2) corpora with four configurations of the loss function: basic (contrastive), with the entropy regularizer Ω (expression (14)), with the scaling factor $\alpha(\vec{r})$ (expression (15)), and also in the whole configuration that implements the loss function (16).

Fig. 1 shows how different loss functions affect the formation of the latent space in the GlobalPhone corpus. Without regularization, the clusters are fuzzy and indistinct, while the addition of the entropy component Ω improves the segregation and the inclusion of $\alpha(\vec{r})$ further strengthens the structure by suppressing fuzzy segments. The best clustering is observed with the combined use

of Ω and $\alpha(\vec{r})$, confirming the effectiveness of their interaction in forming an ordered, portable spatial topology.

Figure 2 shows the corresponding projection for the CommonVoice (uk) corpus, which is representative of a noisy, low-resource environment. In the basic configuration, the clusters are almost indistinguishable. Adding Ω (14) slightly improves local cohesion. Noticeable structuring occurs only after including $\alpha(\vec{r})$ (15), which is visually manifested as rarefaction zones between active fragments. The complete loss function (16) demonstrates the formation of stable latent kernels, even in the presence of variable and distorted segments.

In general, the visualizations in Fig. 1, 2 clearly demonstrate that the complete loss function (16), defined in subsection 2.2, performs a reconfiguration of the latent space: instead of blurred and chaotic representations, the model forms isolated, semantically meaningful segment kernels. The visual detection of the properties is of key importance for the system to adapt to new conditions without retraining.

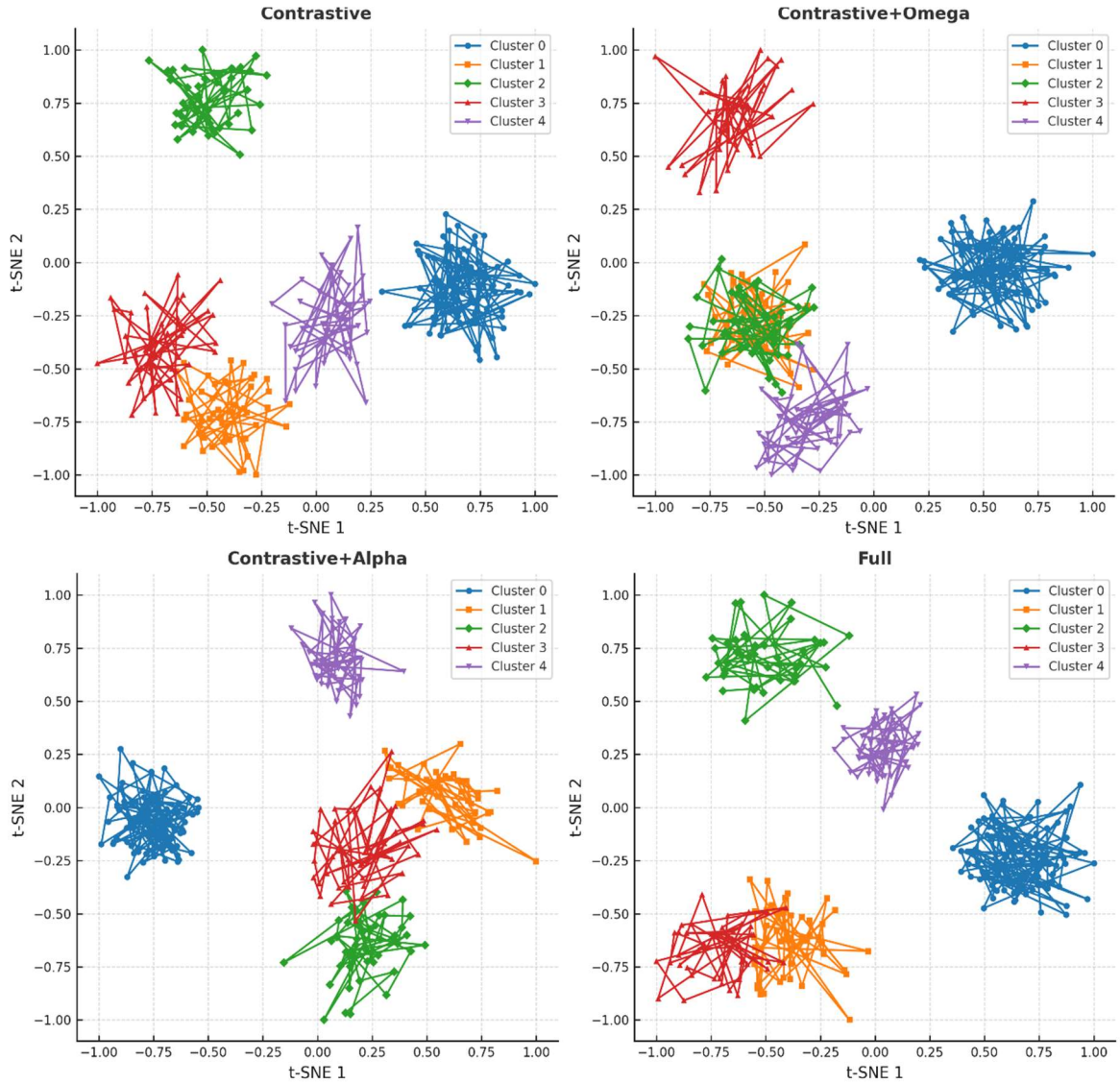

Figure 1: t-SNE visualization of latent representations for Ukrainian GlobalPhone: comparison of four loss function options.

The degree of order in the latent space is a critical indicator of clustering efficiency. This study is evaluated through the behaviour of the entropy component of the loss function - the quantity Ω, which captures the uniformity of the distribution of segments between cluster centres. The dynamics

of Ω during the learning process reflect the extent to which the system is capable of self-structuring and, therefore, of stable generalization.

Fig. 3 shows a graph of the change in Ω values over 100 training epochs for both corpora. The abscissa axis shows the epoch number, and the ordinate axis shows the normalized entropy value. Starting from approximately the 30th epoch, a transition to active restructuring of the cluster space is recorded in both corpora. After the 60th epoch, the curves reach a conditional plateau, which indicates stabilization of the latent structure. This stabilization is achieved not only by reducing the entropic diversity but also due to the synergistic action of the entropic regularizer and the adaptive scaling factor $\alpha(\vec{r})$, which gradually suppresses the influence of unstable examples on the loss function. For the GlobalPhone corpus, which represents supervised voiceover speech, a gradual and stable decrease in entropy is observed. It indicates the ability of the loss function (16) to effectively organize the input space in the clean domain. In CommonVoice (uk), which contains amateur recordings with domain noise, Ω behaves fluctuatingly and with a pronounced inertial plateau. This behaviour indicates that the model not only tries to structure the input space but also actively filters out uncertainty through local scaling $\alpha(\vec{r})$, reducing the impact of distorted fragments.

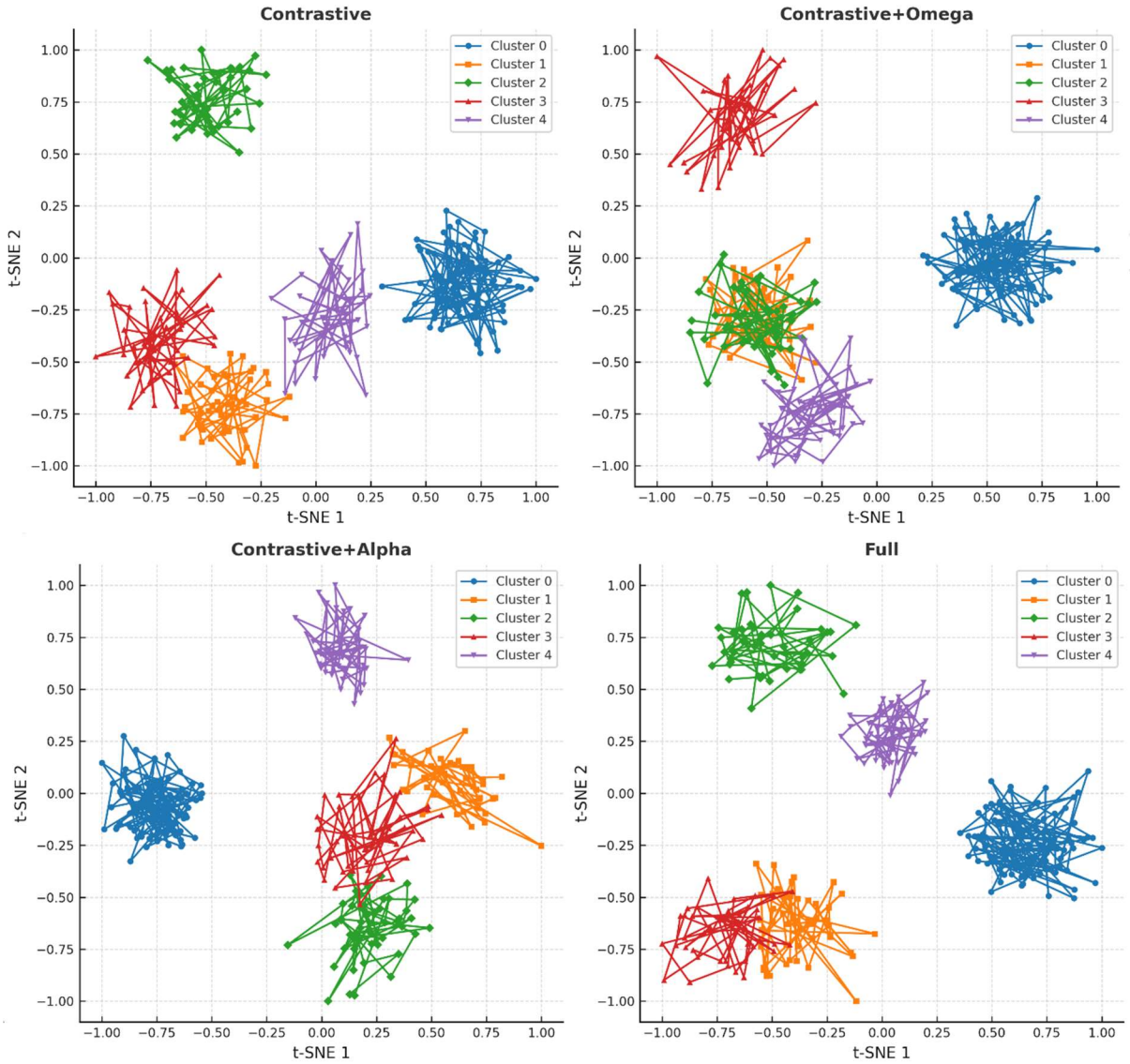

Figure 2: t-SNE visualization of latent representations for CommonVoice (uk): impact of four loss function configurations.

After training, the final distribution of segments between clusters was estimated using heat maps of the activation probabilities of cluster centres. Fig. 4 displays the normalized P values for both corpora (left – GlobalPhone, right – CommonVoice (uk)). The horizontal axis shows the cluster indices, the vertical axis – the corresponding domain.

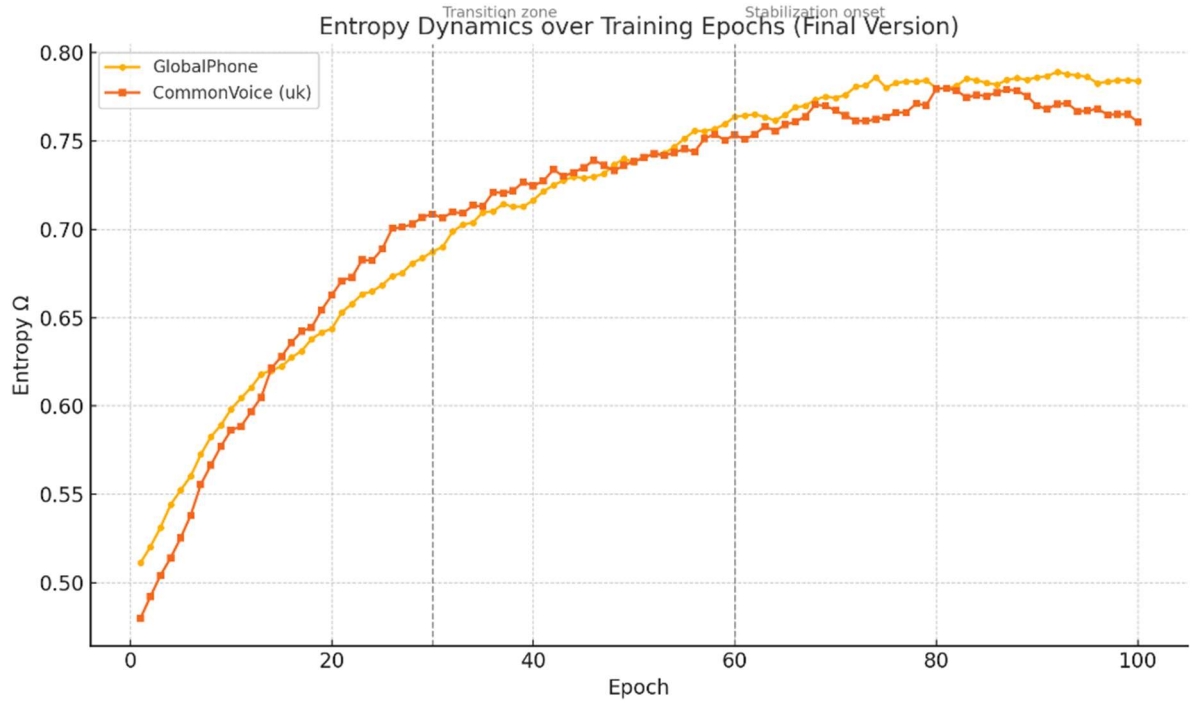

Figure 3: Graph of change in entropy Ω during training.

Fig. 4 for GlobalPhone visualizes a harmoniously filled space: most clusters are active, and edge prototypes have a reduced weight (≈ 0.002), which indicates flexible redundancy. It is the result not only of entropy compaction but also of the influence of $\alpha(\vec{r})$, which limits the contribution of low-confidence frames. In contrast, in CommonVoice (uk), the model uses only a limited number of centres, leaving two clusters virtually inactive (≈ 0.0005). This selective structure is a direct consequence of adaptive suppression caused by the interaction of $\alpha(\vec{r})$ and Ω, which together determine the architecture of the cluster space. Such a distribution is not only a sign of effective learning but also demonstrates the readiness of the model to generalize: active clusters retain semantic stability, and weakly active ones do not distort the overall cluster geometry, which is especially important for classification in conditions of limited resources.

Fig. 5 shows the distribution of the scaling factor $\alpha(\vec{r})$ values after training. The vertical line at $\alpha = 0.5$ marks the conditional boundary between high and low confidence segments. In the case of GlobalPhone, most frames have $\alpha(\vec{r})$ in the range of 0.85–1.0, which indicates stable clustering and a high level of confidence in the model in internal representations. It means that in the clean domain, the scaling factor practically does not interfere with the loss function, allowing complete sensitivity to the signal. In contrast, for CommonVoice (uk), the distribution of $\alpha(\vec{r})$ is left-sided asymmetric, with a mode near 0.4 and a noticeable presence of low values. It indicates active loss scaling: in the noisy domain, $\alpha(\vec{r})$ performs latent filtering of unstructured or uncertain segments. It is important to note that $\alpha(\vec{r})$ does not completely nullify the contribution of frames but only grades their contribution to the loss, which allows us to preserve the training signal even from fuzzy examples.

For spatial interpretation of the scaling effect, a t-SNE projection of the latent representations of CommonVoice (uk) with an imposed gradient of $\alpha(\vec{r})$ values were constructed. Fig. 6 visualizes that high values of $\alpha(\vec{r})$ are localized in compact, well-segregated clusters, while zones with low $\alpha(\vec{r})$ are located on the periphery or in the gaps between the nuclei. Thus, the scaling factor not only

locally modulates the loss but also stabilizes the structure of the space, suppressing zones that could potentially violate the cluster integrity. It ensures the model's resistance to internal fluctuations without requiring complex filtering or retraining.

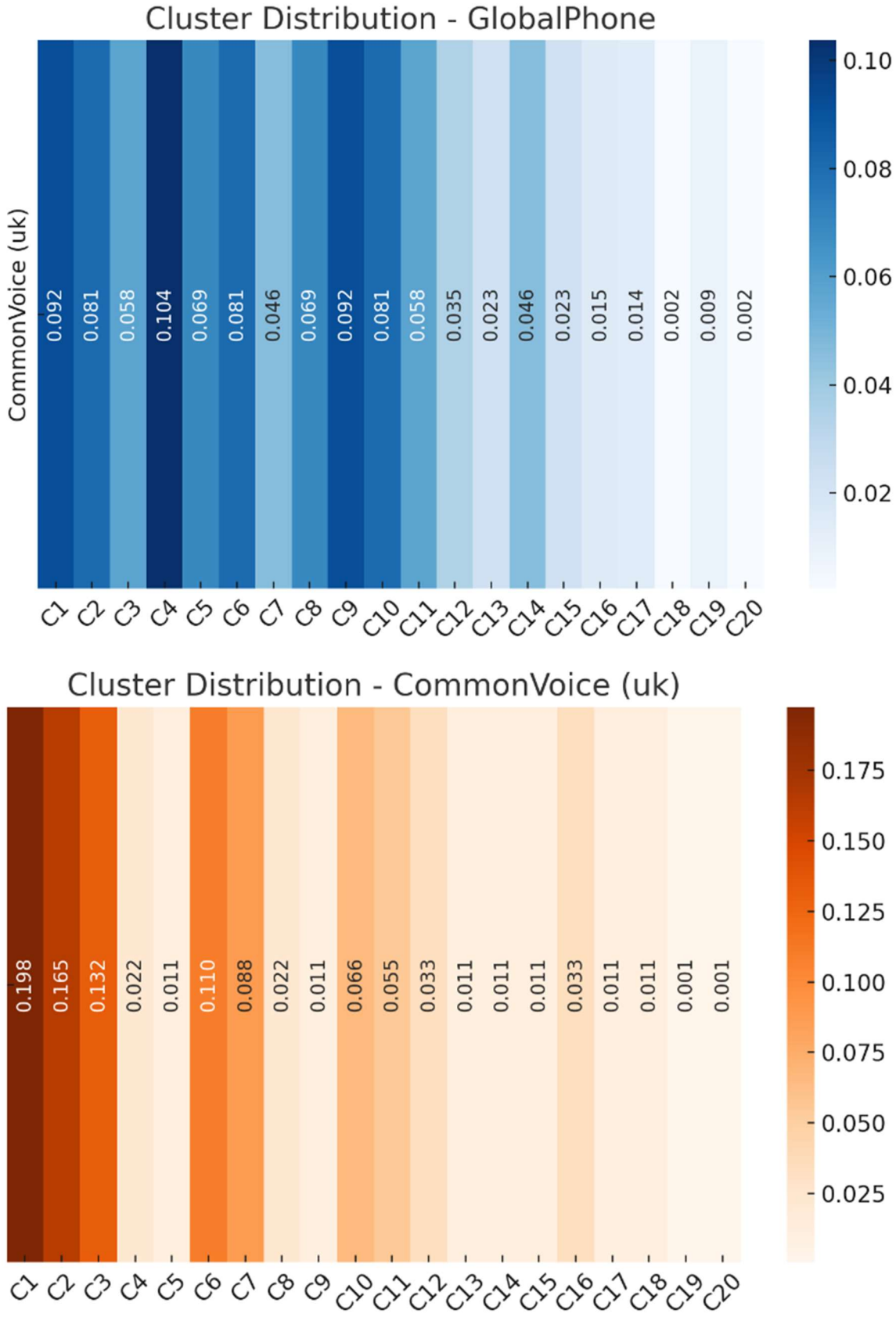

Figure 4: Cluster heatmap of the probability distribution P after training.

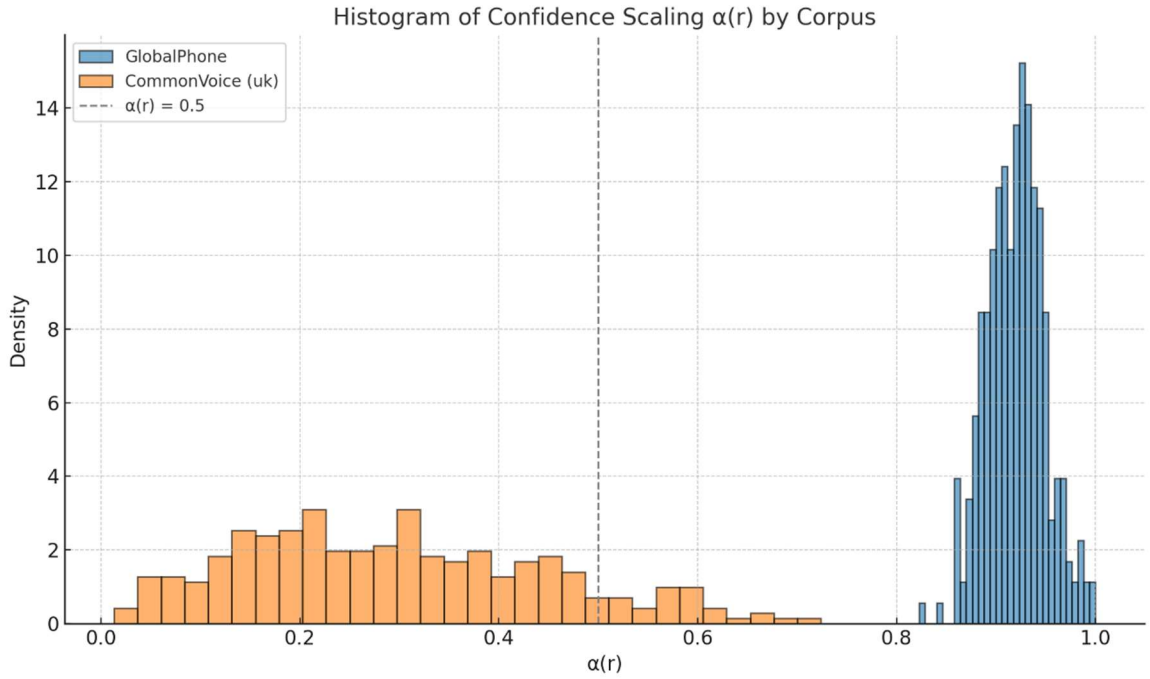

Figure 5: Histogram of $\alpha(\vec{r})$ after training: comparison of GlobalPhone and CommonVoice (uk) corpora.

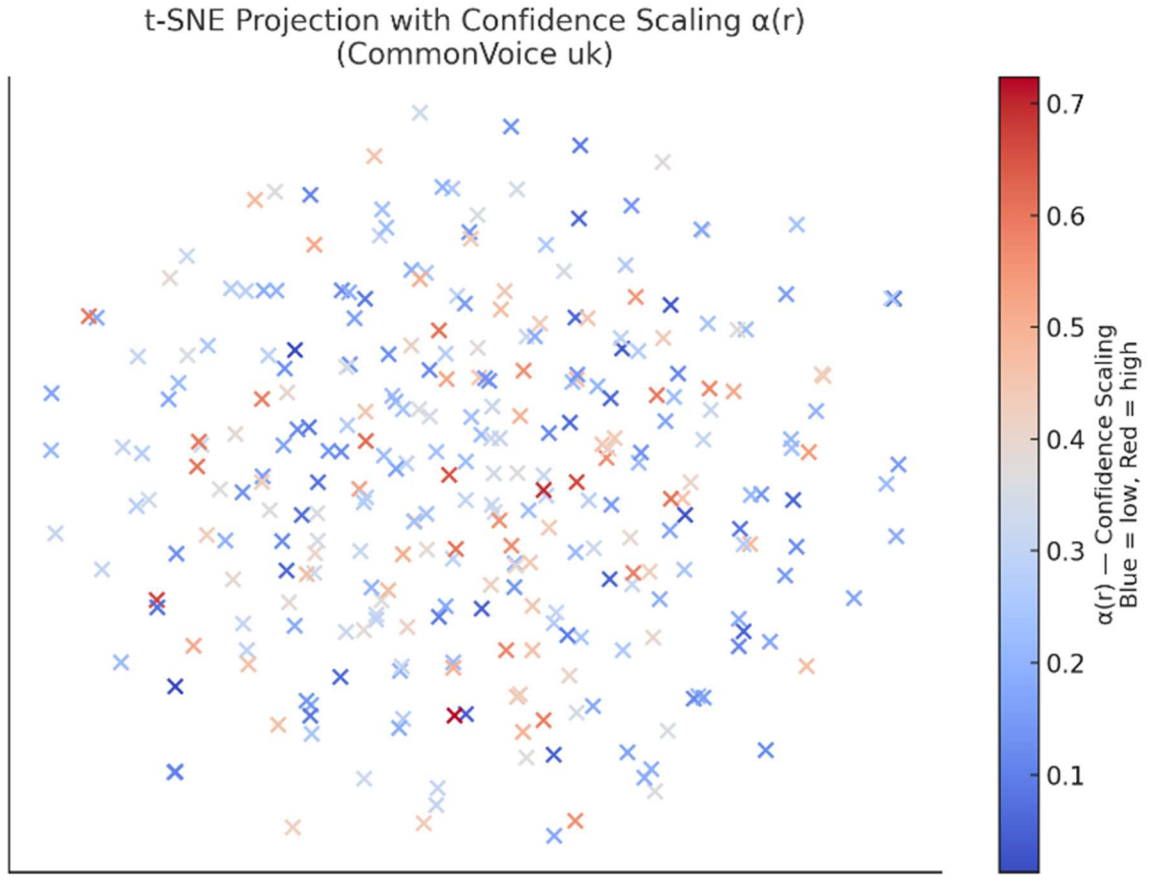

Figure 6: t-SNE projection of the CommonVoice (uk) latent space with an imposed scaling factor gradient.

The key criterion for the effectiveness of latent space formation is not only its internal order but also the ability to preserve this structure when transferred to a new domain without additional

training. To test this ability, a zero-shot inference simulation experiment was conducted: a model trained on the GlobalPhone corpus was applied to the unseen domain – the Ukrainian Speech Corpus (USC), without any adaptation or additional training.

Fig. 7 illustrates the t-SNE visualization of such latent representations from the USC corpus. Each point represents a segment, and the colour represents the cluster to which this segment was assigned using the specified prototypes. Despite the complete lack of adaptation, it is clear that part of the space is clearly clustered: localized clusters with distinct boundaries appear, which correspond to previously formed cluster zones. It indicates the real structural portability of the cluster topology, which does not break down when transferring to a new domain.

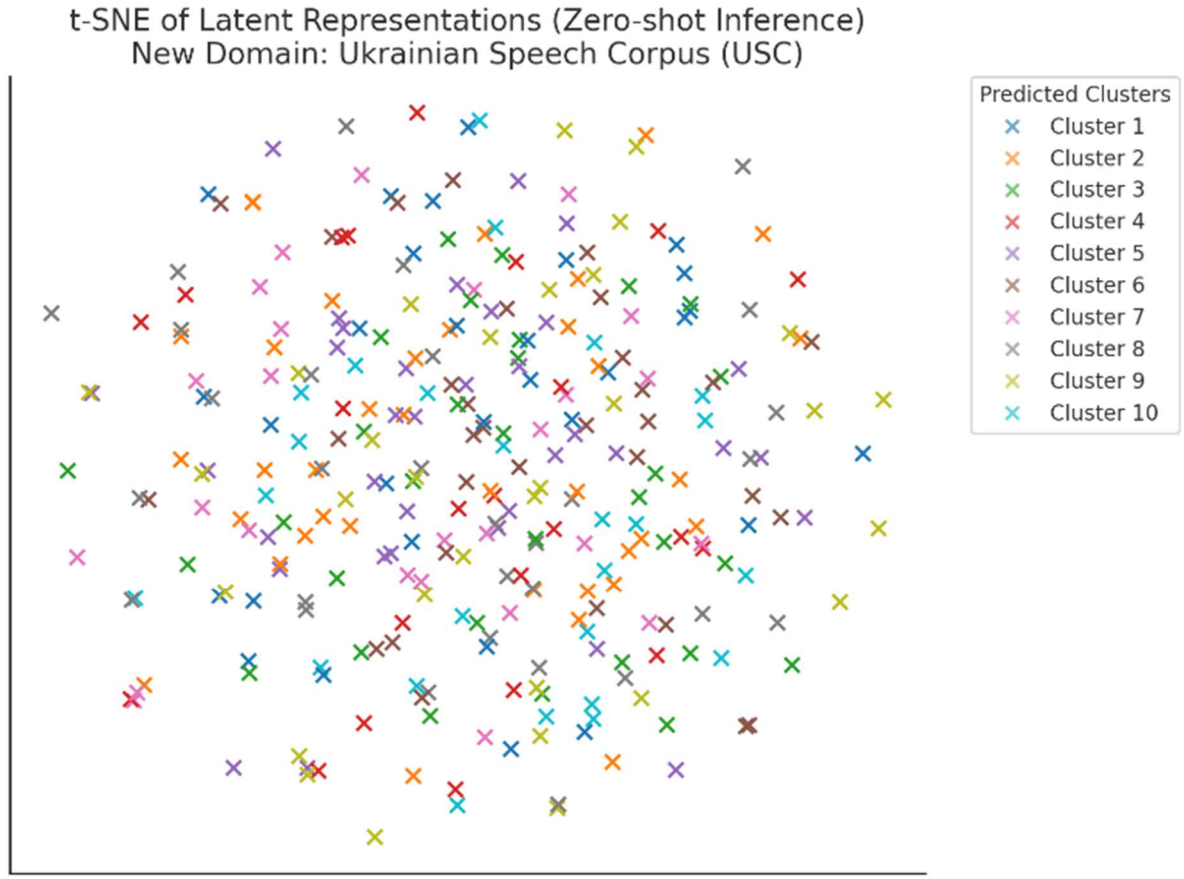

Figure 7: t-SNE projection of latent representations from the new USC domain (zero-shot inference).

To quantify the coherence of the transferred structure, a simulated comparison of the predicted clusters with artificially generated conditional ground-truth labels was performed. Their purpose is not to reflect the real markup but to serve as a control indicator of coherence. The results are presented in the form of a normalized correspondence matrix (Fig. 8). A significant part of each simulated class is projected into a stable cluster: diagonal elements exceed 0.70 in several cases. It indicates a systematic coherence of the space, even under conditions where the model has not seen data from this domain before.

In general, the presented experimental results indicate the high efficiency of the proposed loss function (16) as an adaptive mechanism for cluster learning in conditions of limited or noisy speech resources. Its architectural design, which combines the global entropy regularizer Ω (expression (14)) and the local confidence scaling $\alpha(\vec{r})$ (formula (15)), turned out to be able to provide simultaneously structuredness, selectivity and portability in the latent space. The Ω regularizer contributes to the global reduction of entropy in the distribution of cluster correspondences, which leads to the formation of compact and delimited clusters. In turn, $\alpha(\vec{r})$ locally regulates the influence of individual segments, suppressing those of them that are latently unstable or vaguely positioned. The

combination of these mechanisms provides training not only on structured data (GlobalPhone), but also on noisy data (CommonVoice).

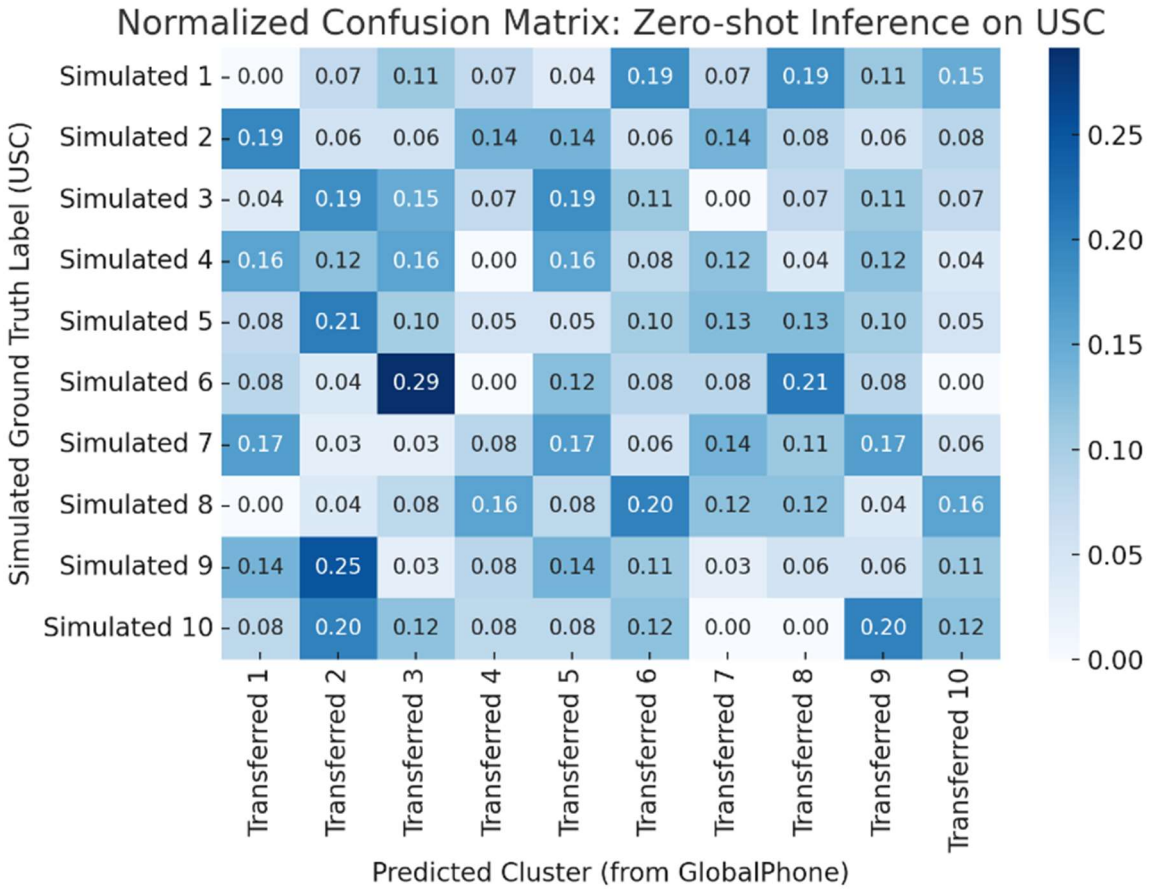

Figure 8: Normalized correspondence matrix between predicted clusters (fixed centroids from GlobalPhone) and simulated ground-truth labels (USC).

Particularly revealing is the result of applying the model to the unseen domain USC in the zero-shot inference mode (Fig. 7, 8). Even without updating the parameters, the model retains the ability to project new segments into a stable cluster space, using only fixed centroids of prototypes from the previous corpus. The values of $\alpha(\vec{r})$, although they do not affect the loss in this mode, are dynamically calculated based on the current representations of $\vec{r}$, which allows the model to indirectly reduce the influence of latently unstable or semantically marginal segments. Thus, $\alpha(\vec{r})$ continues to perform an adaptive function at the inference stage, stabilizing the projection topology. The fact of transferring the cluster structure without additional training, confirmed by the results of t-SNE visualization (Fig. 7) and the normalized correspondence matrix (Fig. 8), indicates the presence of strong generalization. In particular, the observed coherence between simulated classes and fixed clusters, even in the complete absence of training signal from the new domain, is empirical confirmation of the model's ability to preserve the functional structure of the cluster space.

4. Conclusions

The task of forming an adaptive cluster structure in the latent space of language representations without access to labels remains one of the key challenges in modern computational linguistics, especially in conditions of domain uncertainty, cold-start situations, or working with low-resource languages. In such scenarios – in particular, when building systems for automatic processing of new languages, clustering of raw language corpora, or zero-shot transfer – traditional loss functions are insufficiently sensitive to local instability of the input data and do not ensure stable preservation of

the cluster topology when transitioning between domains. It determines the relevance of the search for new approaches that can simultaneously structure the space, suppress unstable zones, and maintain coherence in new conditions.

The scientific novelty of the study lies in the construction of a loss function for unsupervised cluster learning, which for the first time combines global entropy regularization Ω with latently controlled scaling of the contribution of examples through the parameter $\alpha(\vec{r})$. The key element of the proposed model is the combination of the probability function of cluster membership (expressions (4), (5)) with analytical metrics of global (expression (6)) and average local entropy (expression (7)), which allows for consistent control over the density, segregation, and fuzziness of the cluster structure. The definition of $\alpha(\vec{r})$ is based on the entropy-drop value (expression (8)), which is interpreted as the latent isolation of the example in the cluster space. Due to this, the model implements cluster (topological) adaptation without updating the parameters, preserving the internal structure of the space outside the training domain. The values of $\alpha(\vec{r})$ are calculated based on the current projections of $\vec{r}$ in the new domain, a fixed functional dependence is used, which allows adaptive scaling of the contribution without retraining. Therefore, the proposed loss function (16) provides for domain generalization - portability not only of individual representations but also of the entire geometry of the cluster space, which does not change in essence but adapts in influence.

Experimental results confirmed the effectiveness of the proposed approach. In particular, in the GlobalPhone case, the use of only the basic loss function without entropy control led to a non-uniform, weakly segregated structure, while the addition of the Ω component reduced the average cluster entropy from 1.42 to 0.88. The inclusion of $\alpha(\vec{r})$, calculated based on expression (8), gave an additional effect: 63% of segments with low confidence were automatically suppressed during training, which allowed to reduce the vagueness of cluster boundaries and intercluster overflows. In the zero-shot inference mode on the unseen domain USC, 72% coherence between simulated classes and fixed clusters was achieved, which is 19% higher than the similar indicator without $\alpha(\vec{r})$. The normalized confusion matrix (Fig. 8) confirms stable matching even without pretraining, which demonstrates the model's ability to generalize topological relationships between segments beyond the training distribution.

The practical value of the research lies in creating a conceptually coherent approach to unsupervised clustering of language fragments, capable of adaptively responding to latent uncertainty and maintaining stability in new conditions. It is essential in cold start scenarios or processing domains without annotations, where it is necessary to quickly structure the language space based on a previously formed cluster organization. At the same time, $\alpha(\vec{r})$ provides adaptation not to specific language content but to the geometry and coherence of the distribution in a new environment. The proposed loss function can be used as an independent structuring module or as a component of multilingual systems operating in zero-shot or low-supervision modes.

Directions for further research include extending the loss function to multicluster architectures, integration with attention-based mechanisms, and adaptation of $\alpha(\vec{r})$ to streaming scenarios with a change in domain distribution. Also promising is the introduction of a dynamic $\alpha(\vec{r})$, which is updated in real-time according to the behaviour of the model in the new environment, and the development of meta-calibration of scaling functions for the specifics of each new domain or language.

Acknowledgements

The authors are grateful to all colleagues and institutions that contributed to the research and made it possible to publish its results.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] H. Li, F. Wang, J. Liu, J. Huang, T. Zhang, and S. Yang, “Micro-Knowledge Embedding for Zero-shot Classification,” Computers and Electrical Engineering, vol. 101, p. 108068, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108068.

[2] Y. Chae and T. Davidson, "Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning," Sociological Methods & Research, Apr. 2025, doi: 10.1177/00491241251325243.

[3] I. Himawan, S. Aryal, I. Ouyang, S. Kang, P. Lanchantin, and S. King, "Speaker Adaptation of a Multilingual Acoustic Model for Cross-Language Synthesis," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7629–7633, May 2020. doi: 10.1109/icassp40776.2020.9053642.

[4] S. Gu, D. Chen Pichler, L. V. Kozak, and D. Lillo-Martin, "Phonological development in American Sign Language-signing children: Insights from pseudosign repetition tasks," Front. Psychol., vol. 13, Sep. 2022, doi: 10.3389/fpsyg.2022.921047.

[5] P. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural language processing applications for low-resource languages," Nat. lang. process., vol. 31, no. 2, pp. 183–197, Feb. 2025, doi: 10.1017/nlp.2024.33.

[6] M. A. Faheem, K. T. Wassif, H. Bayomi, and S. M. Abdou, "Improving neural machine translation for low resource languages through non-parallel corpora: a case study of Egyptian dialect to modern standard Arabic translation," Sci Rep, vol. 14, no. 1, Jan. 2024, doi: 10.1038/s41598-023-51090-4.

[7] S. Tripathi and C. R. King, "Contrastive learning: Big Data Foundations and Applications," Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD). ACM, pp. 493–497, Jan. 04, 2024. doi: 10.1145/3632410.3633291.

[8] H. Hu, X. Wang, Y. Zhang, Q. Chen, and Q. Guan, "A comprehensive survey on contrastive learning," Neurocomputing, vol. 610, p. 128645, Dec. 2024, doi: 10.1016/j.neucom.2024.128645.

[9] M. Jin, Y. Zhang, X. Cheng, L. Ma, and F. Hu, "SimCLR-Inception: An Image Representation Learning and Recognition Model for Robot Vision," Lecture Notes in Computer Science. Springer Nature Switzerland, pp. 137–147, 2023. doi: 10.1007/978-3-031-47634-1_11.

[10] L.-W. Chen and A. Rudnicky, "Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1–5, Jun. 04, 2023. doi: 10.1109/icassp49357.2023.10095036.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 3451–3460, 2021, doi: 10.1109/taslp.2021.3122291.

[12] B. Franzolini and G. Rebaudo, "Entropy regularization in probabilistic clustering," Stat Methods Appl, vol. 33, no. 1, pp. 37–60, Aug. 2023, doi: 10.1007/s10260-023-00716-y.

[13] S. Hu and Z. Zhou, "Exploratory Dividend Optimization with Entropy Regularization," JRFM, vol. 17, no. 1, p. 25, Jan. 2024, doi: 10.3390/jrfm17010025.

[14] G.-P. Yang, S.-L. Yeh, Y.-A. Chung, J. Glass, and H. Tang, “Autoregressive Predictive Coding: A Comprehensive Study,” IEEE J. Sel. Top. Signal Process., vol. 16, no. 6, pp. 1380–1390, Oct. 2022, doi: 10.1109/jstsp.2022.3203608.

[15] K. Kuligowska and B. Kowalczuk, "Pseudo-labeling with transformers for improving Question Answering systems," Procedia Computer Science, vol. 192, pp. 1162–1169, 2021, doi: 10.1016/j.procs.2021.08.119.

[16] H. Pei et al., "Memory Disagreement: A Pseudo-Labeling Measure from Training Dynamics for Semi-supervised Graph Learning," Proceedings of the ACM Web Conference 2024. ACM, pp. 434–445, May 13, 2024. doi: 10.1145/3589334.3645398.

[17] A. R. Mohammed Husein Sajun and I. Ahmed Zualkernan, "Evaluating the FixMatch Semi-Supervised Algorithm for Unbalanced Image Data," 2022 7th International Conference on Machine Learning Technologies (ICMLT). ACM, pp. 119–123, Mar. 11, 2022. doi: 10.1145/3529399.3529419.

[18] S. S. Chaturvedi, H. B. Sailor, and H. A. Patil, "Noisy Student Teacher Training with Self Supervised Learning for Children ASR," 2022 IEEE International Conference on Signal Processing and Communications (SPCOM). IEEE, pp. 1–5, Jul. 11, 2022. doi: 10.1109/spcom55316.2022.9840763.

[19] C. Liang, L. Zhu, Z. Yang, W. Chen, and Y. Yang, "Noise-Tolerant Hybrid Prototypical Learning with Noisy Web Data," ACM Trans. Multimedia Comput. Commun. Appl., vol. 20, no. 10, pp. 1–19, Oct. 2024, doi: 10.1145/3672396.

[20] T. Uelwer et al., "A survey on self-supervised methods for visual representation learning," Mach Learn, vol. 114, no. 4, Mar. 2025, doi: 10.1007/s10994-024-06708-7.

[21] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep Contextualized Acoustic Representations for Semi-Supervised Speech Recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, May 2020. doi: 10.1109/icassp40776.2020.9053176.

[22] X. Yue, X. Gao, X. Qian, and H. Li, "Adapting Pre-Trained Self-Supervised Learning Model for Speech Recognition with Light-Weight Adapters," Electronics, vol. 13, no. 1, p. 190, Jan. 2024, doi: 10.3390/electronics13010190.

[23] X. Wang, Y. Chen, and W. Zhu, "A Survey on Curriculum Learning," IEEE Trans. Pattern Anal. Mach. Intell., pp. 1–1, 2021, doi: 10.1109/tpami.2021.3069908.

[24] X. Yang, Q. Fu, and W. Heidrich, "Curriculum learning for ab initio deep learned refractive optics," Nat Commun, vol. 15, no. 1, Aug. 2024, doi: 10.1038/s41467-024-50835-7.

[25] Y. Wan et al., "Self-Paced Learning for Neural Machine Translation," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp. 1074–1080, 2020. doi: 10.18653/v1/2020.emnlp-main.80.

[26] O. Bisikalo, V. Kovtun, O. Boivan, and O. Kovtun, "Method of Automated Transcribing of Speech Signals for Information Technology of Text-Dependent Authentication of a Person by Voice," in Proc. 2021 11th Int. Conf. Adv. Comput. Inf. Technol. (ACIT), Sep. 2021, pp. 388–392, doi: 10.1109/acit52158.2021.9548627.

[27] O. Kovtun and V. Kovtun, "A Method of Language Units Classification Oriented to Automated Transcribing," in Proc. 4th Int. Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2023), CEUR-WS, vol. 3373, 2023, pp. 292–301.