pISSN 2005-8063
eISSN 2586-5854
2025. 6. 30.
Vol.17 No.2
pp. 49-57

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.2.049



# Enhancing Matryoshka speaker embeddings with partial element sharing*

Sunchan Park · Hyung Soon Kim**

*Department of Electronics Engineering, Pusan National University, Busan, Korea*

**Abstract**

Matryoshka Representation Learning (MRL) enables efficient extraction of variable-dimensional embeddings from a single high-dimensional vector, offering flexibility in resource-constrained scenarios. However, its strict nested structure, where all elements of lower-dimensional embeddings are shared with higher dimensions, can limit representational power, particularly impacting speaker recognition performance at lower dimensions. To address this limitation, we propose Partial Element Sharing (PES), a technique that enhances Matryoshka speaker embeddings. PES introduces dimension-specific non-shared elements alongside the shared elements inherent in MRL, allowing each embedding dimension to learn more specialized features while maintaining efficiency. Speaker verification experiments on the VoxCeleb dataset demonstrated that PES consistently outperforms standard MRL across various embedding dimensions and evaluation sets. On average, PES achieved up to a 4.9% relative improvement in Equal Error Rate (EER) compared to MRL. Analysis indicates that incorporating non-shared elements improves performance, especially for lower-dimensional embeddings constrained by MRL's structure. PES offers a valuable approach for applications requiring improved speaker recognition performance beyond standard MRL while retaining adaptable dimensionality.

**Keywords:** speaker recognition, speaker verification, speaker embedding, matryoshka representation learning

## 1. 서론

화자 인식은 개인의 음성에 내재된 고유한 특성을 분석하여 그 사람의 신원을 판별하는 기술 분야이다. 화자 인식은 크게 화자 검증, 화자 식별, 화자 분할과 같은 세부 분야로 나눌 수 있다. 화자 검증은 입력된 음성이 특정 등록 화자의 것인지 일대일로 비교하여 화자 일치 여부를 판별한다. 화자 식별은 다수의 등록 화자 집합 내에서 입력된 음성의 화자를 찾아내는 과정이다. 화자 분할은 하나의 오디오 스트림 내에 존재하는 여러 화자의 발화 구간을 분리하고 각 구간에 해당하는 화자를 판별하는 작업이다. 최신 화자 인식 기술은 세부 분야에 관계없이 대부분 화자 임베딩을 기반으로 수행된다. 화자 임베딩은 음성 신호에 포함된 화자 고유 특징을 벡터 공간에 표현한 것으로, 주로 심층 신경망을 이용하여 추출된다. 화자 임베딩의 품질은 전체 시

* This work was supported by a 2-Year Research Grant of Pusan National University.
** kimhs@pusan.ac.kr, Corresponding author
Received 29 April, 2025; Revised 30 May, 2025; Accepted 13 June, 2025

스템의 성능을 결정짓는 핵심 요소로 작용하므로, 고품질의 화자 임베딩을 생성하는 것이 우수한 화자 인식 성능을 달성하는 데 필수적이다. 이를 위해 효과적인 네트워크 구조를 설계하고 목적에 맞는 손실 함수를 선택하는 것이 중요한 연구 주제로 다루어진다.

화자 임베딩 추출을 위한 초기 대표적인 심층 신경망 모델로는 시간 지연 신경망(time-delay neural network, TDNN) 구조 기반의 x-vector가 있다(Snyder et al., 2018). 이후 화자 인식 성능을 개선하기 위해 다양한 TDNN 기반 모델들이 제안되었는데, 특히 ECAPA-TDNN은 TDNN 구조에 다양한 구성 요소들을 효과적으로 통합하여 상당한 성능을 달성하였다(Desplanques et al., 2020). 한편, ResNet과 같은 합성곱 신경망(convolutional neural network, CNN) 기반 구조들도 화자 임베딩 추출에 널리 적용되었다. 이러한 네트워크는 시간-주파수 음향 특징을 이미지처럼 취급하여 처리한다. ResNet 기반 모델들은 다양한 규모에서 전반적으로 좋은 성능을 보여주었다(Wang et al., 2023). 최근에는 TDNN과 CNN 구조를 결합하여 두 방식의 장점을 모두 활용하려는 모델들이 제안되고 있다(Thienpondt & Demuynck, 2023; Yakovlev et al., 2024).

화자 임베딩 학습의 손실 함수는 크게 거리 학습과 분류 기반 방식으로 나뉜다. 거리 학습 기반 손실 함수는 임베딩 간 거리를 직접 최적화할 수 있지만, 효과적인 샘플 선택 전략이 필수적이며 이는 계산 비용 부담이 클 수 있다(Sun et al., 2023; Wang et al., 2019; Zhang et al., 2018). 반면, 분류 기반 손실 함수는 각 학습 샘플이 어떤 화자에 해당하는지를 분류하는 방식으로 학습하므로, 샘플 선택 전략에 따른 계산 비용 부담에서 상대적으로 자유롭다. 하지만 일반적인 분류 기반 손실 함수만으로는 높은 임베딩 품질을 기대하기 어려운데, 이를 보완하기 위해 마진 기반 소프트맥스 손실 함수가 널리 사용된다. 이 방식은 클래스 간 거리는 벌리고 클래스 내 거리는 좁히도록 학습 과정에 마진을 적용하여 임베딩의 변별력과 표현력을 크게 향상시킨다(Han et al., 2023; Liu et al., 2023; Xiang et al., 2019).

이와 더불어 임베딩 자체의 구조와 효율성을 개선하려는 표현 학습 기법에 대한 연구도 활발히 진행되고 있다. 그중 하나가 마트료시카 표현 학습(Matryoshka representation learning, MRL)이다. MRL은 서로 다른 차원의 여러 임베딩을 중첩된 구조를 가진 단일 고차원 벡터 하나로 부호화하는 표현 학습 방법이다. MRL을 사용하면 추론 시 상황에 맞게 추가 비용 없이 다양한 차원의 임베딩을 선택적으로 활용할 수 있다. MRL로 학습된 모델은 최대 차원에서는 해당 차원으로 학습된 기존 모델과 유사한 성능을, 낮은 차원에서는 동일 차원으로 개별 학습된 모델보다 오히려 더 우수한 성능을 나타내는 것으로 보고되었다(Kusupati et al., 2022). 한편 화자 임베딩에도 MRL을 적용한 마트료시카 화자 임베딩이 제안되었는데, 이는 저장 공간과 검색 시간을 크게 줄임으로써 대규모 데이터를 효율적으로 다룰 수 있게 한다. 실험 결과에 따르면 저차원 임베딩에서는 화자 검증 성능이 크게 향상되는 반면, 고차원 임베딩에서는 단일 차원 모델 대비 성능적 한계가 일부 관찰되었다(Wang et al., 2024). 최근 연구에서는 이와 같이 마트료시카 화자 임베딩의 차원에 따라 발생하는 불균형 문제를 완화하기 위한 방법들도 제안되었다(Park & Kim, 2025).

본 연구에서는 마트료시카 화자 임베딩의 성능을 추가로 개선하기 위해 MRL을 확장한 부분 요소 공유(partial element sharing, PES) 기법을 제안한다. 마트료시카 화자 임베딩은 MRL의 구조에 의해 상대적으로 낮은 차원 임베딩의 모든 요소가 상대적으로 높은 차원 임베딩과 공유되는 특징을 갖는다. 이러한 구조는 저장 공간 관점에서는 효율성을 가지지만, 각 차원 임베딩의 표현력을 제한함으로써 화자 인식 성능을 저하시킬 수 있다. PES는 각 차원별 고유 요소를 도입하고, 일부 요소만 임베딩 간에 공유하도록 허용함으로써 이러한 문제를 완화한다. VoxCeleb 데이터셋 기반 화자 검증 실험 결과, PES 적용 시 여러 세부 평가 세트 및 임베딩 차원에 걸쳐 MRL 대비 평균 4.9%의 성능 개선이 달성 가능함을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 방법과 관련된 기존 연구들을 살펴본다. 3장에서는 PES의 세부적인 구조와 학습 방식에 대해 설명한다. 4장에서는 실험에 사용된 데이터셋, 평가 지표, 그리고 비교를 위한 모델 및 방법 등 실험 환경 설정을 다룬다. 5장에서는 실험 결과를 바탕으로 PES의 화자 검증 성능을 MRL 및 단일 차원 모델과 비교 분석한다. 6장에서는 연구 결과를 요약하고 향후 연구 방향을 언급하며 결론을 맺는다.

## 2. 관련 연구

### 2.1. 마트료시카 표현 학습

마트료시카 표현 학습은 고정된 차원의 벡터만 제공하는 기존 표현 학습의 한계를 극복하고, 요구 조건에 따라 다양한 차원의 임베딩을 제공할 수 있도록 설계되었다(Kusupati et al., 2022). 기존에는 연산량 또는 저장 공간의 제약이 발생하는 경우, 별도의 모델을 학습하거나 차원 축소 방법을 통해 더 낮은 차원의 임베딩을 제공해야 했다. 인간이 정보를 다양한 척도로 인식하는 방식에 착안하여, MRL은 단일 고차원 임베딩 안에 여러 수준의 상세 정보를 부호화한다. 이를 통해 단일 임베딩 내의 중첩된 여러 차원의 임베딩들로 다양한 수준의 연산량과 저장 공간 요구 조건에 유연하게 대응할 수 있다. 특히 MRL은 이러한 유연성을 각 차원에 대한 별도의 학습이나 추론 단계에서의 변환 없이, 단일 학습 과정을 통해 효율적으로 확보한다는 장점이 있다.

MRL의 목표는 다음과 같이 나타낼 수 있다. 차원이 $d$인 임베딩 벡터에 대해, $N$을 각 차원 $n \in N$이 $d$보다 작거나 같은 중첩 임베딩 차원의 집합으로 정의한다. MRL은 샘플 데이터 $\mathbf{x}$에 대한 임베딩 $\mathbf{z} \in \mathbb{R}^d$를 학습하는 것을 목표로 하며, 동시에 각 $n \in N$에 대해 $\mathbf{z}$의 첫 $n$개 요소가 임베딩 $\mathbf{z}^{(1:n)} \in \mathbb{R}^n$으로 사용될 수 있도록 한다. 이러한 구조 덕분에 서로 다른 차원의 임베딩을 추가적인 변환 없이 독립적으로 활용할 수 있다. 입력

데이터 $\mathbf{x}$에 대한 임베딩 $\mathbf{z}$는 학습 가능한 파라미터 $\Theta_F$를 가진 심층 신경망 $F$를 사용하여 $\mathbf{z}=F(\mathbf{x};\Theta_F)$와 같이 얻어진다.

MRL의 일반적인 적용 분야인 다중 클래스 분류 문제에 이를 적용하는 경우, 각 차원 $n$에 대한 분류기가 필요하다. 모든 $n$에서의 클래스 수가 $c$로 일정할때, 분류기는 가중치 $\mathbf{W}_n \in \mathbb{R}^{n\times c}$과 편향 $\mathbf{b}\in\mathbb{R}^c$으로 구성될 수 있다. 각 차원 $n$에 대해, 샘플 단위의 소프트맥스 손실 $\ell_{\text{softmax}}$는 $n$차원 임베딩 $\mathbf{z}^{(1:n)}$, 해당 분류기 가중치 $\mathbf{W}_n$, 편향 $\mathbf{b}$, 그리고 샘플 데이터 $\mathbf{x}$의 클래스 레이블 $y$를 사용하여 계산된 후, 상대적 중요도 계수 $c_n \geqq 0$로 가중합된다. 네트워크 파라미터 $\Theta_F$, 그리고 모든 $n$에 대한 분류기 가중치 $\mathbf{W}_n$과 편향 $\mathbf{b}$를 최적화하기 위한 MRL의 샘플 단위 목적 함수는 다음과 같이 정의된다.

$$\ell_{\text{MRL}} = \sum_{n\in N} c_n \cdot \ell_{\text{softmax}}\left(\mathbf{z}^{(1:n)}, y; \mathbf{W}_n, \mathbf{b}\right). \quad (1)$$

한편, MRL에서 각 임베딩 차원마다 필요한 다중 분류기의 메모리 부담을 줄이기 위해, 효율적 마트료시카 표현 학습(efficient Matryoshka representation learning, MRL-E) 기법도 함께 제안되었다. 메모리 소비를 줄이기 위해 MRL-E에서는 모든 분류기에 걸쳐 가중치 공유 메커니즘을 적용한다. 구체적으로, 차원별 가중치 행렬 $\mathbf{W}_n \in \mathbb{R}^{n\times c}$ 대신, MRL-E는 단일 가중치 행렬 $\mathbf{W}\in\mathbb{R}^{d\times c}$을 활용한다. 그리고 각 차원 $n$의 분류기는 전체 가중치 행렬 $\mathbf{W}$의 첫 $n$개 행으로 구성된 부분 행렬 $\mathbf{W}^{(1:n)}\in\mathbb{R}^{n\times c}$를 사용한다. 이러한 가중치 공유 방식은 특히 클래스 수나 임베딩 차원이 클 경우 메모리 요구량을 효과적으로 줄여 학습 효율성을 높이는 데 기여한다. 결과적으로 MRL-E는 MRL의 장점인 다양한 차원의 임베딩 활용 유연성을 그대로 유지하면서도, 메모리 효율성을 개선하고 학습 과정에서의 수렴 속도를 높일 수 있다. MRL-E의 샘플 단위 목적 함수는 다음과 같다.

$$\ell_{\text{MRL-E}} = \sum_{n\in N} c_n \cdot \ell_{\text{softmax}}\left(\mathbf{z}^{(1:n)}, y; \mathbf{W}^{(1:n)}, \mathbf{b}\right). \quad (2)$$

MRL은 다양한 계산량 및 정확도 요구 조건에 맞춰 유연하게 작동하는 적응형 분류에 효과적으로 활용될 수 있다. 기존의 딥러닝 모델들은 고정된 크기의 임베딩을 사용하기 때문에 입력 데이터의 복잡성이나 분류 결과의 신뢰도와 무관하게 항상 동일한 수준의 연산을 수행한다. 반면, MRL은 입력 특성에 따라 임베딩 차원을 동적으로 조절하는 적응형 분류를 가능하게 한다. 이러한 적응형 분류를 구현하기 위해 일반적으로 임계값 기반의 신뢰도 판별 메커니즘이 사용된다. 이 방식에서는 먼저 저차원 임베딩을 사용하여 테스트 샘플의 분류를 시도한다. 이때 분류 결과의 신뢰도가 미리 설정된 임계값을 넘으면 해당 결과를 채택하고, 그렇지 않으면 더 높은 차원의 임베딩을 사용하여 분류 정확도를 높인다. 이 과정은 충분히 신뢰할 만한 분류 결과가 얻어지거나 최대 임베딩 차원을 사용할 때까지 반복된다. 이러한 접근 방식은 높은 분류 정확도를 유지하면서도 평균적인 계산 비용을 크게 절감할 수 있다. 실제로 MRL을 적용한 적응형 분류는 대규모 이미지 분류 벤치마크인 ImageNet-1K 데이터셋에서 기존 방식과 유사한 분류 정확도를 유지하면서도 임베딩 크기를 최대 14배까지 줄이는 결과를 보였다.

또한, MRL은 계산 요구 조건에 따라 임베딩 차원을 동적으로 조정하는 적응형 검색을 효율적으로 수행할 수 있게 한다. 일반적인 적응형 검색 방식은 후보 목록 생성(shortlisting)과 재순위화(re-ranking)의 두 주요 단계로 구성된다. 적응형 검색 시스템은 효율성을 높이기 위해 저차원 임베딩으로 후보를 먼저 탐색한 후, 정확도를 높이기 위해 고차원 임베딩을 사용하여 검색된 후보들의 순위를 재조정한다. 이러한 접근 방식의 구체적인 예시인 퍼널 검색(funnel retrieval)은 MRL에 기반하여 계층적인 후보 생성 전략을 사용한다. 단일 재순위화 단계 대신, 퍼널 검색은 각 단계에서 후보 목록 크기를 절반으로 줄이는 동시에 임베딩 차원을 두 배로 늘려가며 후보 집합을 반복적으로 정제하여 검색 정밀도를 점진적으로 높인다. 이 전략은 단일 단계 검색 방식과 비교했을 때 유사한 검색 정확도를 유지하면서도 이론적으로는 최대 128배, 실제 환경에서는 14배의 속도 향상을 달성하는 것으로 나타났다.

### 2.2. 화자 인식에서의 MRL(Matryoshka representation learning) 적용

마트료시카 화자 임베딩은 MRL을 적용하여 화자 임베딩을 학습함으로써, 단일 임베딩 벡터에서 다양한 차원의 임베딩들을 추출할 수 있다. 마트료시카 화자 임베딩 $\mathbf{z}$에서 추출된 $n$차원 임베딩 $\mathbf{z}^{(1:n)}$은 별도의 계산 과정 없이 화자 유사도 계산에 바로 활용될 수 있다는 장점을 가진다. 학습 과정에서는 각 차원 임베딩의 높은 화자 인식 성능 확보를 위해, 기존 소프트맥스 대신 마진 기반 소프트맥스 손실 함수를 사용한다.

마진 기반 소프트맥스 손실 함수로 학습된 화자 임베딩의 성능은 학습 데이터셋의 전체 화자 수에 영향을 받는데, 일반적으로 학습 데이터에 포함된 화자 수가 많을수록 더 우수한 품질의 화자 임베딩이 생성되는 경향이 있다. 화자 수가 증가하면 분류기 가중치 수도 비례하여 늘어나므로 전체 네트워크가 학습해야 할 파라미터 수가 많아진다. 따라서 학습 데이터의 화자 수가 매우 많은 경우, MRL-E의 가중치 공유 메커니즘이 적용되면 학습 파라미터 수가 줄어들어 마트료시카 화자 임베딩 학습의 효율성을 높이는 데 도움이 된다. MRL은 각 차원의 임베딩마다 별도의 가중치 행렬 $\mathbf{W}_n$을 사용하는 반면, MRL-E는 단일 가중치 행렬 $\mathbf{W}$를 공유하고 각 차원의 임베딩은 $\mathbf{W}$의 특정 부분을 가중치 행렬로 활용한다. MRL-E의 가중치 공유 방식은 하위 임베딩마다 개별 가중치 행렬을 유지할 필요가 없으므로 MRL보다 메모리 사용량이 적다. 또한, 모든 하위 임베딩이 동일한 가중치 행렬을 공유하기 때문에 역전파 과정에서 그래디언트 업데이트가 여러 하위 차원에 동시에 영향을 미쳐 수렴 속도를 높일 수 있다.

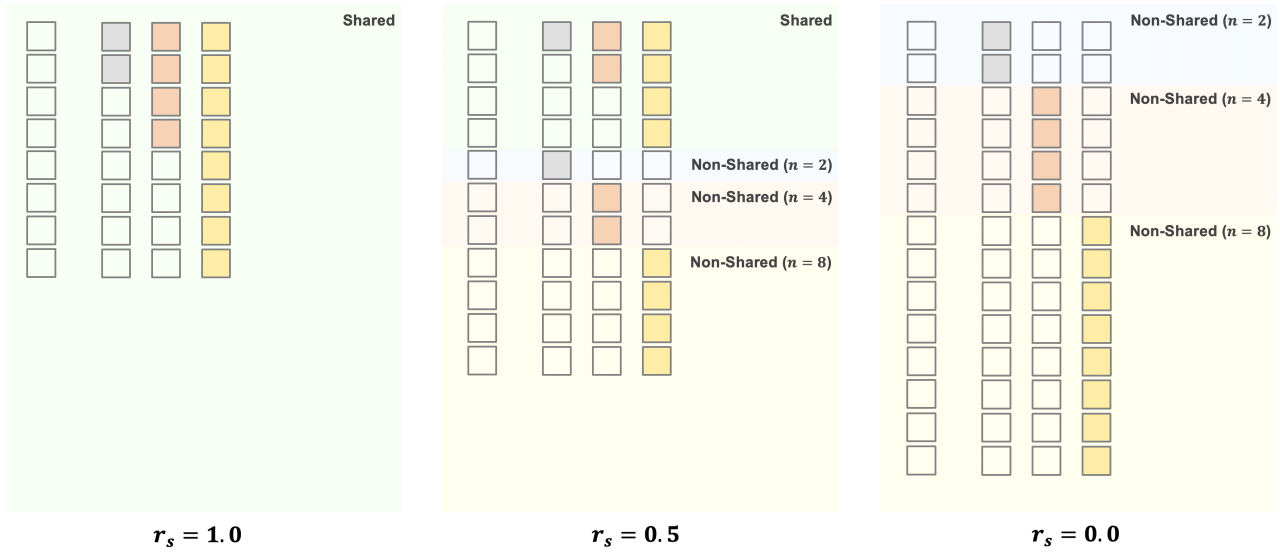


**그림 1.** 부분 요소 공유에 따른 임베딩 구성 예시
**Figure 1.** Example of embedding structures with partial element sharings

마트료시카 화자 임베딩의 주요 장점 중 하나는 매우 낮은 차원에서도 경쟁력 있는 화자 검증 성능을 달성한다는 점이다. 기존 연구의 실험 결과에서 임베딩 차원을 256에서 16으로 줄였을 때 단일 차원 모델의 성능은 9.3배 저하되었지만, MRL 적용 모델은 동일 조건에서 성능 저하 폭이 2.3배에 불과했다. 이 결과는 MRL이 임베딩 차원 감소에 따른 성능 저하를 효과적으로 완화함을 보여준다. 또한 임베딩 차원을 256에서 16으로 줄이면 저장 공간은 94%, 검색 시간은 90% 단축됨을 보고했는데, 이는 마트료시카 화자 임베딩이 낮은 차원을 사용하더라도 높은 효율성과 경쟁력 있는 화자 검증 성능을 동시에 달성할 수 있음을 시사한다. 그러나 높은 차원의 임베딩에서는 이러한 장점이 축소되었는데, 특히 256차원 임베딩에서는 마트료시카 화자 임베딩이 단일 차원 임베딩 대비 오히려 성능이 2.6% 나빠지는 결과를 보여주었다(Wang et al., 2024). 최근 연구에서는 마트료시카 화자 임베딩에서 나타나는 차원별 성능 불균형 문제를 완화하고, 고차원 임베딩의 성능을 개선하는 방법이 제안되었다(Park & Kim, 2025).

## 3. 부분 요소 공유

MRL은 단일 임베딩에서 다양한 차원의 임베딩을 효과적으로 추출하는 기법이다. 그러나 MRL의 구조는 저차원 임베딩이 자신의 모든 요소를 고차원 임베딩과 공유하도록 강제한다. 이러한 엄격한 중첩 구조는 저차원 임베딩일수록 더 많은 상위 차원 임베딩과 요소를 공유하게 만들어, 결과적으로 해당 임베딩의 표현력을 제한하게 된다. 이 때문에 저차원 임베딩은 각 차원에 최적화된 잠재 특징을 효과적으로 학습하는 데 어려움을 겪는다.

본 연구에서는 저차원 임베딩에서 발생 가능한 표현력 제한 문제에 대응하기 위해 부분 요소 공유(PES) 기법을 제안한다. PES는 하위 차원 임베딩의 모든 요소를 상위 차원 임베딩과 공유하는 대신, 각 차원 임베딩별로 지정된 일부 요소만을 공유하도록 허용한다. 이를 통해 각 차원 임베딩은 여러 차원 임베딩들에서 사용되는 공유(shared) 요소와 해당 차원에만 사용되는 비공유(non-shared) 요소를 모두 포함하게 된다. 이는 공유 요소에 의한 제약을 완화시켜 각 차원에 맞는 잠재 특징 학습을 가능하게 하며, 궁극적으로 저차원 임베딩의 표현력이 향상되어 화자 인식 성능 개선에 기여할 수 있다.

각 차원 임베딩에서 공유 요소와 비공유 요소의 수를 조절하기 위해, 공유 비율 $r_s \in [0,1]$를 정의한다. $n$차원 임베딩에서 공유 요소의 수는 $k_n = \lfloor r_s \cdot n \rfloor$으로, 비공유 요소의 수는 $n - k_n$으로 결정된다. 즉, $r_s$가 커질수록 더 많은 요소가 공유되고, 작아질수록 비공유 요소의 비중이 커진다. 특히 $r_s = 1$인 경우 모든 요소가 공유되어 MRL과 동일해지며, $r_s = 0$일 때는 어떠한 요소도 공유되지 않으므로 이 경우를 요소 공유 없음(no element sharing, NES)으로 구분하여 명명한다.

이제 학습하고자 하는 임베딩 차원들의 집합을 오름차순으로 정렬하여 $N = \{n_1, n_2, \cdots, n_{|N|}\}$으로 나타내고, 가장 큰 차원을 $n_{\max}$로 표기한다. 여기서 $|N|$은 집합 $N$의 원소 개수, 즉 고려되는 임베딩 차원의 총 개수를 의미한다. 먼저 $\lfloor r_s \cdot n_{\max} \rfloor$개의 공유 요소들을 할당하고, 각 차원 $n$에 대해 $n - k_n$개의 비공유 요소들을 추가로 설정한다. 전체 임베딩의 차원 $d_p$는 공유 요소의 수 $\lfloor r_s \cdot n_{\max} \rfloor$에 모든 차원 $n \in N$ 각각에 할당된 비공유 요소의 수를 모두 더한 값으로 결정된다.

$$d_p = \lfloor r_s \cdot n_{\max} \rfloor + \sum_{n \in N} (n - k_n). \tag{3}$$

PES의 전체 임베딩 벡터는 여러 요소들이 연결되어 구성되며, 그림 1은 $N = \{2,4,8\}$의 경우에 해당하는 임베딩 구성을 시각적으로 보여준다. 먼저, 공유 요소들의 벡터는 $\mathbf{s}$로 나타내며,

여기에 포함된 요소들은 서로 다른 차원 임베딩 간에 공유될 수 있음을 의미한다. 다음으로, 각 $n$차원 임베딩에서만 사용되는 비공유 요소들의 벡터는 $\mathbf{p}_n$으로 표현한다. PES의 전체 임베딩 $\mathbf{z}$는 공유 요소 벡터 $\mathbf{s}$와 각 차원 $n$에 대한 비공유 요소 벡터 $\mathbf{p}_n$들을 순차적으로 연결(concatenation)하여 다음과 같이 구성된다.

$$\mathbf{z} = [\mathbf{s}, \mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_{|N|}]. \quad (4)$$

한편, PES의 각 $n$차원 임베딩은 공유 요소 벡터 $\mathbf{s}$의 일부와 해당 차원의 비공유 요소 벡터 $\mathbf{p}_n$으로 형성된다. 공유 요소는 앞에서 정의한 것처럼 $\mathbf{s}$의 첫 $k_n$개 요소가 사용되므로, $n$차원 임베딩은 연결을 통해 $[\mathbf{s}^{(1:k_n)}, \mathbf{p}_n]$과 같은 형태로 만들어진다. 이를 바탕으로 PES의 샘플 단위 목적 함수는 다음과 같이 나타낼 수 있다.

$$\ell_{\text{PES}} = \sum_{n \in N} c_n \cdot \ell_{\text{softmax}}\left([\mathbf{s}^{(1:k_n)}, \mathbf{p}_n], y; \mathbf{W}_n, \mathbf{b}\right). \quad (5)$$

추가적으로, 부분 요소 공유에 가중치 공유를 적용한 효율적 부분 요소 공유(efficient partial element sharing, PES-E) 기법을 제안한다. 이 방식에서는 MRL-E와 동일하게 각 차원별 가중치 행렬 $\mathbf{W}_n$ 대신 단일 가중치 행렬 $\mathbf{W}$를 사용하며, $\mathbf{W}$의 첫 $n$개 행으로 구성된 부분 행렬 $\mathbf{W}^{(1:n)}$을 $n$차원 임베딩 학습에 사용한다. PES-E의 샘플 단위 목적 함수는 다음과 같다.

$$\ell_{\text{PES-E}} = \sum_{n \in N} c_n \cdot \ell_{\text{softmax}}\left([\mathbf{s}^{(1:k_n)}, \mathbf{p}_n], y; \mathbf{W}^{(1:n)}, \mathbf{b}\right). \quad (6)$$

## 4. 실험 설정

### 4.1. 데이터셋

본 연구에서는 화자 검증 성능 평가를 위해 VoxCeleb1(Nagrani et al., 2017) 및 VoxCeleb2(Chung et al., 2018) 데이터셋을 사용하였다. 이들은 다양한 환경에서 녹화된 유튜브(YouTube) 인터뷰 영상으로부터 수집되었는데, 인터넷에 공개된 영상으로부터 자동 수집되었기 때문에, 명시적 동의 절차를 거치지 않아 개인정보 보호 문제를 야기할 수 있다. 그럼에도 불구하고, VoxCeleb 데이터셋은 현재 공개적으로 이용 가능한 가장 큰 규모의 화자 인식 데이터셋 중 하나이며, 관련 연구 분야에서 성능 비교를 위한 표준 벤치마크로 널리 사용된다. 따라서 본 연구에서도 기존 연구들과의 공정한 비교를 위해 이 데이터셋을 채택하였다.

화자 검증 연구의 표준적인 실험 설정에 따라, 학습에는 VoxCeleb2의 학습용 데이터를, 평가에는 VoxCeleb1 데이터로 구성된 평가 세트인 VoxCeleb1-O, VoxCeleb1-E, VoxCeleb1-H를 사용하였다. 5,994명 화자와 1,092,009개 발화로 구성된 VoxCeleb2의 학습용 데이터는 화자 임베딩 학습을 위해 가장 널리 사용되는 대규모 데이터셋 중 하나이다. 세 가지 평가 세트 중 가장 작은 VoxCeleb1-O는 40명의 화자로부터 추출된 37,611개의 발화 쌍으로 구성된다. VoxCeleb1-E는 1,251명의 화자로 구성된 VoxCeleb1 데이터 전체에서 581,480개의 발화 쌍을 임의 추출하여 구성된다. 1,190명 화자의 552,536개 발화 쌍으로 구성된 VoxCeleb1-H는 더 어려운 평가를 위해 동일한 성별과 국적을 가진 화자들의 발화쌍을 샘플링하여 구성된다. 이러한 평가 세트들은 다양한 난이도와 화자 구성 조건에서 모델의 성능을 평가하기 위한 표준 벤치마크로서, 기존 화자 인식 연구에서 활용되어 왔다.

### 4.2. 학습 설정

제안하는 방법의 성능 평가를 위해, 화자 임베딩 추출 신경망은 ResNet34 기반 구조로 고정하고 임베딩 차원 및 손실 함수를 변경하며 모델 학습을 수행하였다. ResNet34는 연산 효율이 높으면서도 좋은 성능을 제공하는 모델로, 다양한 설정에 대한 학습 및 평가를 효율적으로 수행할 수 있어 선정되었다. 베이스라인으로 사용된 ResNet34 모델 구조 및 학습 과정은 wespeaker 툴킷(Wang et al., 2023)의 레시피를 바탕으로 설정되었다. ResNet34 모델은 학습 가능한 파라미터로 이루어진 총 34개 레이어로 구성된다. 신경망은 32 채널의 3×3 합성곱 레이어(convolutional layer)와 배치 정규화(batch normalization)로 시작하여, 4단계의 잔차 블록(residual block)들로 연결된다. 이 블록들은 32 채널 블록 3개, 64 채널 블록 4개, 128 채널 블록 6개, 256 채널 블록 3개로 구성된다. 잔차 블록은 두 개의 연속적인 3x3 합성곱 레이어와 배치 정규화로 구성되며, 첫 번째 합성곱 뒤에 ReLU 활성화 함수가, 블록 입력과 출력 사이에는 항등 연결(identity shortcut)이 적용된다. 다운샘플링은 단계 2부터 각 단계의 첫 블록에서 스트라이드(stride) 2 합성곱을 통해 이루어지고, 이후 블록은 스트라이드 1을 사용한다. 프레임 전체의 평균과 표준편차를 모으는 시간 통계 풀링(temporal statistics pooling)을 거쳐, 최종 완전 연결 레이어(fully-connected layer)에서 고정 차원의 화자 임베딩이 생성된다. 베이스라인 모델의 임베딩 차원 $d$는 256, MRL과 PES를 위한 임베딩 차원은 $N = \{16, 32, 64, 128, 256\}$으로 설정하였다.

학습 데이터 준비 과정에는 속도 변화(speed perturbation), 부가 잡음(additive noise), 잔향(reverberation)을 포함하는 데이터 증강 기법이 활용되었다. 속도 변화는 0.9, 1.0, 1.1 중 하나를 동일 확률로 무작위 선택하여 해당 비율로 대상 발화의 속도를 조정하였다. 기존 연구의 설정에 따라, 각 비율의 속도 변화가 적용된 음성은 별도의 화자로 레이블링되었다(Yamamoto et al., 2019). 잡음 및 잔향은 각각 0.6의 확률로 적용되었으며, 잡음은 MUSAN(Snyder et al., 2015) 데이터셋에서, 잔향을 위한 실내 임펄스 응답은 실내 임펄스 응답 및 잡음 데이터베이스(Ko et al., 2017)로부터 확보되었다. 입력 음향 특징(acoustic feature)은 10 ms 프레임 시프트(frame shift)에 25 ms 프레임 윈도우(frame window)로 추출된 80차원 로그 멜 필터뱅크 에너지(log mel-filter bank energy)를 사용하였다. 학습을 위해 각 발화의 음향 특징에

서 200 프레임의 세그먼트를 무작위 추출하고, 각 세그먼트에 평균 정규화(mean normalization)를 적용하였다. 모델 학습을 위한 미니배치(mini-batch)는 서로 다른 발화에서 샘플링된 256개의 세그먼트로 구성되었다.

베이스라인 ResNet34 모델 학습에는 확률적 경사 하강법(stochastic gradient descent, SGD) 옵티마이저와 가산 각도 마진 소프트맥스(additive angular margin softmax, AAM-Softmax) 손실함수가 사용되었다. SGD 옵티마이저의 모멘텀(momentum)은 0.9로, 가중치 감쇠(weight decay)는 0.0001로 설정하였다. 한편 학습률(learning rate)과 AAM-Softmax의 마진 값은 별도의 스케줄러에 의해 제어되었다. 학습률은 총 150 에포크(epoch)의 각 학습 단계(step)에서 0.1에서 0.00005로 지수적으로 감소했으며, 처음 6 에포크에는 0에서 1까지 선형 증가하는 워밍업(warm-up) 계수가 곱해졌다. 마진 값의 경우 초기 20 에포크 동안에는 0, 이후 40 에포크에 이르기까지 0.2를 목표 값으로 지수적으로 증가한 뒤, 40 에포크 이후에는 0.2를 유지되도록 설정되었다. 한편 모든 실험에서 AAM-Softmax의 스케일(scale) 값은 32로, 각 차원의 상대적 중요도 계수 $c_n$은 1로 고정되었다.

### 4.3. 평가 방법

화자 검증 성능을 평가하기 위해, 평가 쌍의 각 등록 및 테스트 발화에서 화자 임베딩을 추출한다. 이후 임베딩 사이 코사인 유사도를 계산하고, AS-Norm(adaptive s-normalization) 기반의 유사도 정규화를 적용한다(Karam et al., 2011). AS-Norm을 위한 임포스터 세트(imposter set) 구성을 위해 학습 데이터에서 화자 임베딩을 추출하고, 평균을 통해 각 화자당 하나의 임베딩을 생성한다. 이 중에서 등록 및 테스트 임베딩 각각과 가장 유사한 상위 300개의 임베딩을 선택하여 유사도 정규화에 사용한다. MRL과 PES에서는 각 차원 임베딩에 대해 위 평가 과정을 반복 수행한다.

평가 지표로는 동일 오류율(equal error rate, EER)을 사용한다. EER은 시스템의 임계값(threshold) 설정에 따라 달라지는 오수락률(false acceptance rate, FAR)과 오거부율(false rejection rate, FRR)이 서로 같아지는 지점에서의 오류율을 나타낸다. EER은 이 두 가지 유형의 오류를 균형있게 반영하는 단일 대푯값으로서 화자 검증 연구에서 널리 사용된다. 관련 연구에서는 최소 탐지 비용 함수(minimum detection cost function, minDCF)와 같은 다른 지표들도 활용되지만, 본 연구에서는 다양한 임베딩 차원의 성능을 여러 평가지표로 비교할 경우 분석이 지나치게 복잡해질 수 있으므로 해당 추가 지표들은 실험 결과에 포함하지 않았다.

## 5. 실험 결과

### 5.1. 전체 실험 결과 및 분석

표 1은 VoxCeleb1-O, VoxCeleb1-E, VoxCeleb1-H 각각에 대해 16, 32, 64, 128, 256차원 임베딩으로 화자 검증 평가를 수행한 결과를 EER로 보여준다. 단일 차원 모델(single-dimensional model, SDM)과 MRL, MRL-E, NES 모델에 대한 평가 결과를 베이스라인으로, 공유 비율 $r_s$가 0.25, 0.5, 0.75인 경우에 대한 PES 및 PES-E 모델에 대한 평가 결과를 비교하였다. MRL과 MRL-E는 각각 PES와 PES-E에서 $r_s$를 1로 설정한 경우와, NES는 $r_s$를 0으로 설정한 경우와 동일하게 동작하므로, 표 1에서 이에 따라 $r_s$ 값을 표기하였다. 표 1에서 SDM의 결과는 각 목표 임베딩 차원마다 독립적인 신경망 모델을 개별적으로 학습 및 평가하고 이를 종합한 것인 반면, 다른 모델들의 결과는 단일 모델 학습 후 각 목표 차원의 임베딩을 추출하여 평가한 것이다.

SDM과 비교했을 때, MRL 및 MRL-E 모델은 낮은 차원에서는 더 우수한 성능을 보였지만, 차원이 높아짐에 따라 성능 차이가 감소하여 256차원에서는 오히려 SDM보다 낮은 성능을 기록했다. 임베딩 차원 증가에 따른 성능 저하 현상은 기존 연구에서도 관찰된 바 있으며, 본 연구의 실험 결과에서도 일관된 경향이 확인되었다. MRL과 MRL-E 모델 간의 성능 차이는 평가 세트 및 차원에 따라 다르지만, 평균 EER을 기준으로 할 때

**표 1.** 각 임베딩 차원에서의 화자 검증 결과(EER, %)
**Table 1.** Speaker verification results across each embedding dimension (EER, %)

| Model | $r_s$ | VoxCeleb1-O | | | | | VoxCeleb1-E | | | | | VoxCeleb1-H | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 16-D | 32-D | 64-D | 128-D | 256-D | 16-D | 32-D | 64-D | 128-D | 256-D | 16-D | 32-D | 64-D | 128-D | 256-D | |
| SDM | | 3.18 | 1.53 | 0.94 | 0.85 | 0.80 | 3.21 | 1.54 | 1.09 | 0.97 | 0.95 | 5.80 | 2.72 | 1.92 | 1.75 | 1.73 | 1.93 |
| MRL | 1.00 | 2.52 | 1.47 | 1.18 | 0.96 | 0.96 | 2.62 | 1.52 | 1.20 | 1.07 | 1.07 | 4.61 | 2.69 | 2.12 | 1.91 | 1.91 | 1.85 |
| MRL-E | 1.00 | 2.40 | 1.39 | 0.99 | 0.94 | 0.94 | 2.58 | 1.53 | 1.15 | 1.09 | 1.09 | 4.63 | 2.68 | 2.05 | 1.96 | 1.96 | 1.83 |
| PES | 0.75 | 2.20 | 1.38 | 1.04 | 0.89 | 0.89 | 2.58 | 1.50 | 1.17 | 1.08 | 1.08 | 4.68 | 2.65 | 2.08 | 1.93 | 1.93 | 1.81 |
| | 0.50 | 2.17 | 1.29 | 1.03 | 0.99 | 0.98 | 2.56 | 1.50 | 1.17 | 1.05 | 1.05 | 4.62 | 2.64 | 2.03 | 1.88 | 1.88 | 1.79 |
| | 0.25 | 2.37 | 1.24 | 0.97 | 0.89 | 0.87 | 2.54 | 1.47 | 1.11 | 1.03 | 1.02 | 4.55 | 2.59 | 2.03 | 1.88 | 1.87 | 1.76 |
| PES-E | 0.75 | 2.48 | 1.52 | 1.05 | 0.97 | 0.97 | 2.53 | 1.51 | 1.15 | 1.08 | 1.08 | 4.58 | 2.66 | 2.03 | 1.92 | 1.92 | 1.83 |
| | 0.50 | 2.13 | 1.44 | 1.01 | 0.95 | 0.95 | 2.51 | 1.52 | 1.14 | 1.10 | 1.10 | 4.51 | 2.61 | 2.04 | 1.97 | 1.97 | 1.80 |
| | 0.25 | 2.12 | 1.32 | 1.06 | 0.92 | 0.92 | 2.51 | 1.49 | 1.14 | 1.05 | 1.05 | 4.56 | 2.64 | 2.04 | 1.91 | 1.91 | 1.78 |
| NES | 0.00 | 2.17 | 1.31 | 1.00 | 0.89 | 0.90 | 2.54 | 1.46 | 1.12 | 1.03 | 1.02 | 4.59 | 2.56 | 2.00 | 1.87 | 1.86 | 1.75 |

EER, equal error rate; SDM, single-dimensional model; MRL, Matryoshka representation learning; PES, partial element sharing; NES, no element sharing.

MRL-E가 MRL 대비 소폭 향상된 성능을 보였다.

PES와 PES-E의 경우 평가 세트 및 차원에 따라 복합적인 패턴이 나타나지만, 평균적으로 $r_s$가 낮아질수록 성능이 향상되는 경향성을 보인다. 이는 더 많은 수의 비공유 요소를 할당할수록 각 차원 임베딩의 표현력이 개선되어, 화자 검증 성능이 전반적으로  향상된 것으로 볼 수 있다. 그러나 $r_s$ 감소에 따른 PES와 PES-E의 성능 향상에도 불구하고, NES가 최고 성능을 나타냈다. 이는 모델 파라미터 공유의 장점을 활용하면서 임베딩 표현에 제약이 없어 각 차원 임베딩의 표현력이 최대화되었기 때문으로 보인다. 하지만 이러한 성능은 더 큰 전체 임베딩 크기를 대가로 한다. 본 실험 설정에서 NES의 전체 임베딩 차원은 496차원에 달하는 반면, MRL은 256차원에 불과하다. 제안하는 PES는 공유 비율 $r_s$에 따라 그 사이의 절충점을 제공하며, 이는 한정된 자원 내에서 성능을 최대한 높여야 하는 응용 환경에서 중요한 의미를 갖는다.

MRL과 MRL-E 사이 성능 차이를 고려할 때, PES는 PES-E 대비 평균 성능이 소폭 앞서는 모습을 보여준다. 이는 PES-E의 가중치 공유 메커니즘이 공유 요소들에만 실질적으로 적용되므로, 비공유 요소의 비중이 높은 설정에서는 그 효과가 상대적으로 감소하기 때문으로 보인다. 모든 평가 세트 및 차원에 대한 평균 EER을 기준으로, PES의 경우 MRL 대비 최대 4.9%, PES-E의 경우 MRL-E 대비 최대 2.7%의 성능 개선 효과를 확인하였다.

**표 2.** PES 및 PES-E 모델의 오류 감소율(%)
**Table 2.** Error reduction rates of PES and PES-E models (%)

| Model | $r_s$ | 16-D | 32-D | 64-D | 128-D | 256-D |
|---|---|---|---|---|---|---|
| PES (vs. MRL) | 0.75 | 3.0 | 2.6 | 4.7 | 1.0 | 1.0 |
| | 0.50 | 4.1 | 4.4 | 6.2 | 0.5 | 0.8 |
| | 0.25 | 3.0 | 6.7 | 8.7 | 3.6 | 4.6 |
| PES-E (vs. MRL-E) | 0.75 | 0.2 | -1.6 | -1.0 | 0.5 | 0.5 |
| | 0.50 | 4.8 | 0.5 | 0.0 | -0.8 | -0.8 |
| | 0.25 | 4.4 | 2.7 | -1.2 | 2.8 | 2.8 |

PES, partial element sharing; MRL, Matryoshka representation learning.

### 5.2. 임베딩 차원별 성능 변화 분석

표 2는 MRL 대비 PES, MRL-E 대비 PES-E의 평균 EER 기준 오류 감소율을 정리한 결과를 보여준다. PES의 경우 $r_s$가 0.75, 0.5인 설정에서 저차원 임베딩(16, 32, 64차원)일 때 고차원 임베딩(128, 256차원)보다 성능 향상 폭이 더 크다. 이는 저차원 임베딩에서 공유 요소들이 더 많은 임베딩들과 공유되므로, 비공유 요소의 도입이 화자 임베딩 성능 향상에 효과적임을 보여준다. 반면, $r_s$가 0.25인 경우에는 고차원 임베딩에서도 성능 향상 효과가 나타났다. PES에서는 모든 실험 설정에서 다른 차원 대비 64차원 임베딩에서의 성능 향상 폭이 가장 크게 나타난다.

PES-E에서 $r_s$가 0.75인 경우, MRL-E 대비 오히려 성능이 저하되는 경우가 발생했다. 이는 가중치 공유의 이점이 적용되는 공유 요소의 수가 줄어든 반면, 추가된 비공유 요소로 인한 성능 향상 효과가 이를 상쇄하지 못했기 때문으로 분석된다. $r_s$가 0.5일 때는 16차원에서 큰 폭의 성능 개선 효과가 나타났으며, 0.25일 때는 64차원을 제외한 모든 차원에서 성능이 개선되었다. 이러한 결과들을 종합하면, PES-E에서도 비공유 요소 비중 증가가 성능 향상에 기여함을 확인할 수 있었다. 한편, PES에서는 64차원에서 가장 큰 성능 향상을 보인 반면, PES-E에서는 64차원에서 성능이 개선되지 않았다. 이는 MRL-E 및 PES-E의 가중치 공유 방식이 64차원과 같은 중간 차원에서 특히 효과적일 수 있음을 시사하며, PES-E에서 비공유 요소 도입으로 인한 추가적인 성능 개선 여지가 상대적으로 제한되었을 수 있음을 의미한다.

## 6. 결론

본 연구에서는 기존 마트료시카 표현 학습(MRL) 기반 화자 임베딩의 추가적인 성능 향상을 위한 부분 요소 공유(PES) 기법을 제안하였다. MRL은 단일 고차원 임베딩에서 다양한 차원의 하위 차원 임베딩을 효율적으로 추출할 수 있는 장점이 있지만, 하위 차원 임베딩의 모든 요소가 상위 차원 임베딩과 공유되는 구조적 제약으로 인해 특히 저차원 임베딩의 표현력이 제한될 수 있다. 제안된 PES 기법은 각 임베딩 차원마다 할당된 비공유 요소와 다른 차원과 공유되는 요소를 함께 사용함으로써 이러한 제약을 완화하고자 하였다. VoxCeleb 데이터셋을 이용한 화자 검증 실험 결과, 제안된 PES 기법은 기존 MRL 대비 전반적으로 향상된 성능을 보였다. 모든 평가 조건에 대한 평균 EER 기준, PES는 MRL 대비 최대 4.9%의 성능 개선을 달성하였다. 또한, 차원별 임베딩 분석을 통해 비공유 요소 도입이 특히 MRL의 제약이 크게 작용하는 저차원 임베딩 성능 향상에 효과적이며, 비공유 요소 비중을 높이면 고차원 임베딩 성능 개선에도 기여함을 확인하였다. 제안된 PES 기법은 저장 공간에 어느 정도 제약이 있지만 MRL보다 우수한 화자 인식 성능이 필요한 응용 환경에서 유용하게 활용될 수 있을 것으로 기대된다. 이는 최고 성능을 보이는 NES 기법에 비해 저장 공간 요구사항을 줄이면서도 성능 저하를 최소화할 수 있어, 성능과 효율성 간의 균형을 맞추는 데 실질적인 이점을 제공한다. 향후 연구에서는 다양한 공유 비율의 임베딩을 얻을 수 있는 신경망 구조를 설계하고, 학습 과정에서 임의의 공유 비율을 샘플링하여 학습하는 방법을 제안하고자 한다. 이를 통해  공유 비율에 따라 별도의 모델을 학습하는 대신, 단일 모델로부터 다양한 공유 비율의 임베딩을 효율적으로 얻을 수 있는 학습 기법 및 구조를 개발할 예정이다. 아울러 각 차원별 손실을 개별 계산하여 합산하는 기존 방식 대신, 학습 데이터셋 화자들에 대한 로그 확률 값 단계에서 통합하는 방법 등 다양한 수준에서 서로 다른 차원 임베딩의 학습을 통합하는 기법에 대한 연구도 진행할 계획이다.

## References

Chung, J. S., Nagrani, A., & Zisserman, A. (2018, September). VoxCeleb2: Deep speaker recognition. *Proceedings of Interspeech*

*2018*, (pp. 1086-1090). Hyderabad, India.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020, October). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification. *Proceedings of Interspeech 2020*, (pp. 3830-3834). Shanghai, China.

Han, B., Chen, Z., & Qian, Y. (2023, June). Exploring binary classification loss for speaker verification. *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 1-5). Rhodes Island, Greece.

Karam, Z. N., Campbell, W. M., & Dehak, N. (2011, May). Towards reduced false-alarms using cohorts. *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 4512-4515). Prague, Czech.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017, March). A study on data augmentation of reverberant speech for robust speech recognition. *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 5220-5224). New Orleans, LA.

Kusupati, A., Rege A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., ... Farhadi, A. (2022, December). Matryoshka representation learning. *Proceedings of Advances in Neural Information Processing Systems 36*, (pp. 30233-30249). New Orleans, LA.

Liu, Q., Zhang, X., Liang, X., Qian, Y., & Yao, S. (2023). AWLloss: Speaker verification based on the quality and difficulty of speech. *IEEE Signal Processing Letters*, *30*, 1337-1341.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017, August). VoxCeleb: A large-scale speaker identification dataset. *Proceedings of Interspeech 2017*, (pp. 2616-2620). Stockholm, Sweden.

Park, S., & Kim, H. S. (2025). Dimension-specific margins and element-wise gradient scaling for enhanced Matryoshka speaker embedding. *IEEE Access*, *13*, 45473-45487.

Snyder, D., Chen, G., & Povey, D. (2015). *MUSAN: A music, speech, and noise corpus.* arXiv. https://arxiv.org/abs/1510.08484.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust DNN embeddings for speaker recognition. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 5329-5333). Calgary, Canada.

Sun, Y., Zhang, H., Wang, L., Lee, K. A., Liu, M., & Dang, J. (2023, June). Noise-disentanglement metric learning for robust speaker verification. *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 1-5). Rhodes Island, Greece.

Thienpondt, J., & Demuynck, K. (2023, December). ECAPA2: A hybrid neural network architecture and training strategy for robust speaker embeddings. *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop*, (pp. 1-8). Taipei, Taiwan.

Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., Deng, Y., & Qian, Y. (2023, June). Wespeaker: A research and production oriented speaker embedding learning toolkit. *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. Rhodes Island, Greece.

Wang, J., Wang, K. C., Law, M. T., Rudzicz, F., & Brudno, M. (2019, May). Centroid-based deep metric learning for speaker recognition. *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 3652-3656). Brighton, UK.

Wang, S., Zhu, P., & Li, H. (2024). *M-Vec: Matryoshka speaker embeddings with flexible dimensions.* arXiv. https://arxiv.org/abs/2409.15782

Xiang, X., Wang, S., Huang, H., Qian, Y., & Yu, K. (2019, November). Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. *Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, (pp. 1652-1656). Lanzhou, China.

Yakovlev, I., Makarov, R., Balykin, A., Malov, P., Okhotnikov, A., & Torgashov, N. (2024, September). Reshape dimensions network for speaker recognition. *Proceedings of Interspeech 2024*, (pp. 3235-3239). Kos Island, Greece.

Yamamoto, H., Lee, K. A., Okabe, K., & Koshinaka, T. (2019, September). Speaker augmentation and bandwidth extension for deep speaker embedding. *Proceedings of Interspeech 2019*, (pp. 406-410). Graz, Austria.

Zhang, C., Koishida, K., & Hansen, J. H. L. (2018). Text-independent speaker verification based on triplet convolutional neural network embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(9), 1633-1644.

• **박순찬 (Sunchan Park)**
부산대학교 전자공학과 박사과정
부산시 금정구 부산대학로63번길 2
Tel: 051-510-1704
Email: sunchanpark@pusan.ac.kr
관심분야: 음성 인식, 화자 인식, 음성 감정 인식

• **김형순 (Hyung Soon Kim)** 교신저자
부산대학교 전자공학과 교수
부산시 금정구 부산대학로63번길 2
Tel: 051-510-2452
Email: kimhs@pusan.ac.kr
관심분야: 음성 인식 및 합성, 화자 인식, 음성 신호 처리

# 부분 요소 공유를 통한 마트료시카 화자 임베딩 향상*

박 순 찬 · 김 형 순

부산대학교 전자공학과

**국문초록**

마트료시카 표현 학습(Matryoshka representation learning, MRL)은 단일 고차원 벡터로부터 가변 차원 임베딩의 효율적인 추출을 가능하게 하여, 자원이 제한된 시나리오에서 유연성을 제공한다. 그러나 하위 차원 임베딩의 모든 요소가 상위 차원과 공유되는 엄격한 중첩 구조는 표현력을 제한할 수 있으며, 특히 낮은 차원에서 화자 인식 성능에 영향을 미친다. 이러한 한계를 해결하기 위해, 본 연구는 마트료시카 화자 임베딩을 향상시키는 기법인 부분 요소 공유(partial element sharing, PES)를 제안한다. PES는 MRL에 내재된 공유 요소와 함께 차원별 비공유 요소를 도입하여, 각 임베딩 차원이 효율성을 유지하면서 더 특화된 특징을 학습하도록 허용한다. VoxCeleb 데이터셋에서의 화자 검증 실험은 PES가 다양한 임베딩 차원 및 평가 세트에 걸쳐 표준 MRL보다 일관되게 우수한 성능을 보임을 입증했다. 평균적으로 PES는 MRL 대비 동일 오류율(equal error Rate, EER)에서 최대 4.9%의 상대적 개선을 달성했다. 분석 결과, 비공유 요소의 통합은 특히 MRL 구조에 의해 제약을 받는 저차원 임베딩의 성능을 향상시키는 것으로 나타난다. PES는 적응 가능한 차원성을 유지하면서 표준 MRL 이상의 향상된 화자 인식 성능을 요구하는 응용 분야에 유용한 접근 방식을 제공한다.

**핵심어:** 화자 인식, 화자 검증, 화자 임베딩, 마트료시카 표현 학습

* 이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.