---

**Proceedings Paper:**
Meghanani, A. and Hain, T. orcid.org/0000-0003-0939-3464 (2024) Deriving translational acoustic sub-word embeddings. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) Proceedings. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 16-20 Dec 2023, Taipei, Taiwan. Institute of Electrical and Electronics Engineers (IEEE) . ISBN 9798350306903

https://doi.org/10.1109/ASRU57964.2023.10389747

---

# DERIVING TRANSLATIONAL ACOUSTIC SUB-WORD EMBEDDINGS

*Amit Meghanani, Thomas Hain*

Speech and Hearing Research Group
Department of Computer Science, The University of Sheffield, United Kingdom
{ameghanani1,t.hain}@sheffield.ac.uk

## ABSTRACT

There is a growing interest in understanding the representational geometry of acoustic word embeddings (AWEs), which are fixed-dimensional representations of spoken words. However, not much research has been conducted on acoustic sub-word embeddings (ASWEs), which can provide a better understanding of the AWE space. This work focuses on decomposing AWEs to obtain ASWEs while retaining the ability to reconstruct AWEs by translating ASWEs in the embedding space, under constrained settings. Initially, high-quality AWEs are obtained with an Average Precision (AP) score of 0.97 on the word discrimination task. Subsequently, ASWEs are derived through the decomposition of AWEs. Three adapted versions of the AP metric, utilized for evaluating the quality of the derived ASWEs and their translational properties, are proposed. The results demonstrate that the derived ASWEs exhibit high quality, and the reconstruction of AWEs from the ASWEs is achievable by translating them in the embedding space.

***Index Terms*—** acoustic word embeddings, acoustic sub-word embeddings, translational, word discrimination task.

## 1. INTRODUCTION

Deep learning methods have evolved considerably over the last decade and the use of embeddings are prevalent across many tasks as a way to encapsulate and express a wide range of properties of inputs. In natural language processing, embeddings are often used to represent discrete objects in semantic space, as in the work Word2vec [1] and BERT [2]. In speech processing, embeddings are typically used to encode signal context at regular time intervals as in Wav2vec [3] and HuBERT [4] models or for segments of speech such as i-vector [5] and X-vector [6]. In addition, there has been a long-standing interest in the linguistic unit (word and sub-word) embeddings for the purpose of unit discovery [7, 8, 9, 10] and acoustic word embeddings (AWE) for word matching [11].

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. This work was also funded in part by LivePerson, Inc.

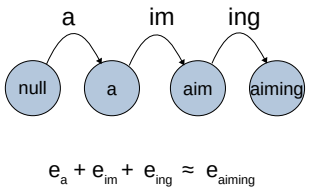


**Fig. 1**. Demonstration of the evolving linguistic units (words or sub-words) by translating them in embedding space with acoustic sub-word embedding vectors. The figure shows the proposed hypothesis of reconstructing AWE of a word "aiming" by translating its ASWEs.

Acoustic word embeddings are attractive in processing because they are fixed length representations of variable length speech signal that encodes acoustic-phonetic content rather than pure signal information.

Recently, AWEs have been studied in greater detail, in terms of extraction and use [12, 13, 14, 15]. Also, there is a growing interest in analyzing the representational geometry of AWEs [16] and its ability to capture phonological similarity [17]. AWEs are also getting attention in cognitive science as it exhibits a word onset bias, which is reported in various studies on human speech processing and lexical access [13]. Nevertheless, not much has been reported on finding acoustic sub-word embeddings (ASWEs). This work attempts to find ASWEs by decomposing AWEs in the embedding space, under constrained settings. The AWEs are decomposed in such a way that they can be reconstructed back by using the derived ASWEs. We hypothesize that the reconstruction process can be described by simple translations in embedding space, as shown in Fig. 1. The translations are performed by ASWEs, and in this case, the translations are simple addition operations in the embedding space. From Fig. 1, it can be seen that the word-like unit "aim" can be interpreted as the sub-word-like embedding of "a" plus the sub-word-like embedding of "im". Further, the word-like unit "aiming" can be interpreted as the word-like embedding of "aim" plus sub-word-like embedding of "ing". This hypothesis is inspired by TransE [18] work. TransE models hierarchical relationships by interpreting them as vector translations operating on the

embeddings of the entities. For example, if $(h, r, t)$ is a triple from a knowledge graph where $h, t \in E$ (a set of entities) are two entities with relationship $r$ ($h, r, t \in \mathbb{R}^d$), then $h + r \approx t$ if the given fact is true else $h + r \neq t$. Similarly, sub-word embeddings can be understood as translations: the addition of a sub-word unit to an existing linguistic unit (word or sub-word) can be interpreted as translations in embedding space determined by a sub-word embedding as shown in Fig. 1.

In summary, position-dependent translational ASWEs are obtained by decomposing AWEs, under constrained settings. These ASWEs represent the compositional model of AWEs. The classes of the sub-words are defined based on the byte pair encoding (BPE) tokens [19] learned from the text modality. Additionally, by using learned ASWEs, AWEs of "written words" are also obtained (Sec 4.2, Sec 5), even for unseen words during training. Therefore, "spoken words' or "written words" can be characterized by these acoustic sub-word embeddings. These understandings can be applied to various downstream tasks. For example, a written word or text can be project into an acoustic embedding space for audio-text agreements. Another application would be zero-shot open vocabulary keyword spotting, where keywords are enrolled with text-only, without any spoken examples [20]. The main contributions of this work are as follows:

1. This work is the first attempt to derive translational ASWEs, under constrained settings, which models the relationship between linguistic units (sub-words or words) as simple translations in the embedding space as demonstrated in Fig. 1.
2. Using AWEs only, we describe an efficient method for obtaining translational BPE token-based ASWEs without using sub-word time boundary information. Reconstruction of AWEs by translating ASWEs in the embedding space is also demonstrated.
3. Three adapted versions of AP (used for word discrimination task to evaluate AWEs) are proposed to evaluate ASWEs and their translational properties (Sec. 3.2).
4. A method to obtain acoustic word embeddings from text modality (**Word2AWE**) is presented. High quality AWEs of "written words" (even for unseen words during training) are constructed from the ASWEs (Sec. 4.2, Sec. 5) by using their translational properties.

The rest of the paper structure is as follows: Sec. 2 and 3 describes proposed methodology and implementation details for extracting AWEs and ASWEs, respectively; Sec. 4 and 5 describes the experiments and results; Sec. 6 concludes the work with potential future directions.

## 2. PROPOSED METHODOLOGY

The extraction process of ASWE is divided into two steps. The first step is to compute AWEs and the second step is to extract ASWEs by decomposing AWEs. Both steps are described in the following subsections:

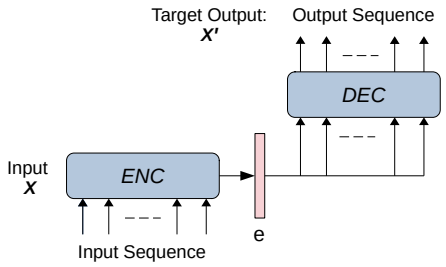


**Fig. 2**. CAE-RNN model setup for AWE [12].

### 2.1. Acoustic Word Embedding Extraction

AWEs are extracted using correspondence autoencoder-recurrent neural networks (CAE-RNN) [21], which shows promising performance on the acoustic word discrimination task [12]. CAE-RNN is trained with the pairs of segments having different instances of the same spoken word $(X, X')$. $X = x_1, x_2..., x_T$ and $X' = x'_1, x'_2..., x'_{T'}$ are sequences of observed acoustic feature vectors extracted from the speech segments. Fig. 2 illustrates the CAE-RNN model used in our experiments. After processing the acoustic input vectors $X$, the encoder produces the acoustic word embedding $e$ as shown in Fig. 2 and Eq. 1.

$$e = ENC(X) \tag{1}$$

This embedding is then fed to the decoder as input at every time step [21], whose target output is $X'$. The mean squared loss for a single training pair $(X, X')$ is shown in Eq. 2.

$$L(X, X') = \sum_{t=1}^{T'} ||x'_t - f_t(X)||^2 \tag{2}$$

where $f_t(X)$ is the $t^{\text{th}}$ decoder output.

By feeding two different instances of the same word to CAE-RNN, it is ensured that the generated embeddings are invariant to irrelevant properties (e.g. speaker, channel, duration) whilst at the same time capturing the spoken word identity. MFCC [22] and HuBERT [4] features are explored as inputs to train CAE-RNN model.

### 2.2. Acoustic Sub-word Embedding Extraction

The aim of the work is to validate the proposed translational acoustic sub-word embedding hypothesis and establish a proof of concept. Therefore, as a first experiment towards this research direction, the dataset is defined as a collection of words with a duration greater than or equal to 0.5 seconds (standard choice in the literature) [23] and three sub-words.

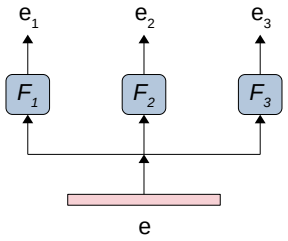


**Fig. 3**. Decomposition model for extracting ASWEs. The model has three feedforward neural networks for sub-word classification with joint training. The ASWEs $(e_1, e_2, e_3)$ are extracted from the last layer of the networks.

More details about data preparation are given in Sec. 3.1. The model used for decomposing AWEs into ASWEs is shown in Fig. 3.

The decomposition model has three parallel feedforward neural networks (FNN) ($F_1, F_2, F_3$, no shared parameters), one for each position-dependent sub-word, jointly trained as a sub-word classifier. AWEs are the inputs for this model as shown in Fig. 3. For each network, the output of the last hidden layer is taken as a sub-word embedding representation as shown in Eq. 3, where $e_1, e_2$, and $e_3$ are ASWEs of sub-word at positions 1, 2, and 3 in a word, respectively. The total loss of the model is the sum of three individual losses as shown in Eq. 4, where $L_{e_1}, L_{e_2}$ and $L_{e_3}$ are the cross-entropy losses of a single training instance for each FNN to learn position-dependent ASWEs. However, to ensure the translational property between the extracted ASWEs $(e_1, e_2, e_3)$, as described in Sec. 1, the similarity loss term is added to the objective loss function as shown in Eq. 5. Similarity loss is calculated between the reconstructed embedding $e'$ by translating the ASWEs ($\approx e_1 + e_2 + e_3$) and the original input embedding $e$. The reconstructed embedding $e'$ is calculated as mentioned in Eq. 6, where $F_{proj}$ is a fully connected projection layer. Experiments are conducted on all variations of loss represented as $L_a$ and $L_b$ in Eq. 4, 5, respectively with two variations of reconstructed embedding $e'$ as shown in Eq. 6.

$$e_i = F_i(e) \quad \text{for} \quad \text{i = 1, 2, 3} \tag{3}$$

$$L_a = L_{e_1} + L_{e_2} + L_{e_3} \tag{4}$$

$$L_b = L_a + \underbrace{(1 - \cos(e', e))}_{\text{similarity loss}} \tag{5}$$

where $e'$ is defined as:

$$e' = \begin{cases} F_{proj}(e_1 + e_2 + e_2), & \text{If projected} \\ e_1 + e_2 + e_3, & \text{otherwise} \end{cases} \tag{6}$$

**Table 1**. A summary of word and sub-word statistics in train, validation, and test splits.

| Dataset split | Train | Validation | Test |
|---|---|---|---|
| **Unique spoken words** | 5152 | 3399 | 3360 |
| **Total spoken words** | 29334 | 9778 | 9778 |
| **Unique spoken sub-words** | 152 | 152 | 152 |
| **Total spoken sub-words** | $29334 \times 3$ | $9778 \times 3$ | $9778 \times 3$ |

**Algorithm 1** : Data sampling method

**Input:** $f(w)$: Relative frequency histogram of words; $f(s)$: Relative frequency histogram of sub-words; $S = (s_1, s_2..., s_n)$ are unique sub-words and $Z = (z_{s_1}, z_{s_2}, ..., z_{s_n})$ are their relative frequencies; N = maximum unique words to be sampled (= 6000 in this case)
**Output:** $D$ : Sampled dataset
$Z' = (1 - z_{s_1}, ..., 1 - z_{s_n})$ ▷ inverse relative frequency
$M \leftarrow 0$ ▷ unique words in the sampled data
$D = \{\}$
**while** $M \leq N$ **do**
  $s_i \leftarrow f(s), Z'$ ▷ sample $s_i$ from $f(s)$ with weights $Z'$
  $w_j \leftarrow f(w)$ ▷ sample $w_j$: least frequent word having $s_i$
  $D \leftarrow D + w_j$
  Recompute $f(w), f(s), Z'$
  **if** $w_j \in D$ **then**
    Continue
  **else**
    $M \leftarrow M + 1$ ▷ update M only when $w_j \notin D$
  **end if**
**end while**

## 3. IMPLEMENTATION DETAILS

### 3.1. Data Preparation

Experiments are conducted on the LibriSpeech dataset [24]. The LibriSpeech dataset is force-aligned to obtain the boundaries of word segments. Then, a byte pair encoding (BPE) [19] tokeniser is trained with a vocabulary size of 200 on available text from LibriSpeech recordings. A proof of concept is formulated to validate the proposed hypothesis of obtaining translational ASWEs by decomposing AWEs in embedding space. The dataset is defined as a collection of words with 3 BPE tokens (sub-words) and a duration greater than or equal to 0.5 seconds (a standard choice in the literature) [23]. The sub-words obtained are imbalanced due to the presence of both rare and frequent sub-words. To mitigate this issue, a simple sampling method (Algorithm 1) is applied. Following sampling, the dataset comprises 6,000 unique words, a reduced set of 152 sub-words, and a total of 48,890 instances of spoken words with time boundaries. The dataset is divided into training, validation, and test sets by the ratio of 60%, 20%, and 20%, respectively. Table 1 displays the total spoken words and unique words available in the training, validation, and test splits. The code and dataset are available on GitHub[1]

[1]https://github.com/Trikaldarshi/ASWE.git

### 3.2. Proposed Metrics

The average precision (AP) metric for the word discrimination task is used in the literature to evaluate AWEs [25, 26, 27, 28, 29]. It was first proposed in the work [30] for rapid evaluation of speech representations for spoken term discovery. To evaluate ASWEs, three adapted versions of this AP metric used for the word discrimination task are proposed here.

For the acoustic word discrimination task, a pair of AWEs are compared to decide whether they belong to the same word or not. For the acoustic word discrimination task, all possible spoken word pairs are generated. If there are $M$ spoken words, then the total number of generated spoken word pairs will be $\binom{M}{2} = \frac{M(M-1)}{2}$. For each pair, cosine distance between their AWEs is computed. Then, the performance measure is the AP, which is the area under the precision-recall curve generated by varying all possible thresholds and has a maximum value of 1. The proposed metrics for ASWEs are as follows:

1. *AP-SD (Average Precision - Sub-word Discrimination):* Average precision on the sub-word discrimination task for each position, denoted as AP@$p_1$, AP@$p_2$, and AP@$p_3$. It is an adapted version of the acoustic word discrimination task, where AWEs will be replaced by ASWEs.

2. *AP-RW (Average Precision - Reconstructed Words) :* Average precision on the acoustic word discrimination task between original and reconstructed AWEs. AP-RW is useful for evaluating translational properties of the learned ASWEs as AWEs are reconstructed by translating ASWEs in embedding space. If $\mathcal{W} = \{w_i\}_{i=1}^{M}$ represents the original spoken words and $\mathcal{W}' = \{w_i'\}_{i=1}^{K}$ represents the reconstructed AWEs then the total number of possible spoken word pairs will be $(w_i, w_j') \in \mathcal{W} \times \mathcal{W}'$ (total $M \times K$ pairs). Then the AP will be calculated as described earlier.

3. *AP-CW (Average Precision - Constructed Words) :* Average precision on the acoustic word discrimination task between original AWE and the constructed AWE of a "written word". This metric is useful for measuring the quality of the constructed AWEs of "written" words. The spoken word pairs will be generated from constructed AWEs of a "written word" and original AWEs. If there are M original spoken words and $K$ constructed AWEs of "written" words, then the total number of possible spoken word pairs will be $M \times K$. Then the AP will be calculated as described earlier.

### 3.3. AWE Implementation Details

To train the CAE-RNN model, a total of 2,57,557 pairs of spoken word segments having different instances $(X, X')$ are generated from the training set. To extract features of a spoken word, first, the features for the entire sentence are obtained. Then, the available time segments of the spoken word are used to extract its features. MFCC [22] and HuBERT [4] are explored as input acoustic features extracted from the spoken word segments. For each spoken word, 20-dimensional MFCC features are extracted with 30 ms window size and 20 ms shift along with delta and delta-delta features, which provide 60-dimensional MFCC feature vectors. For HuBERT features, 768-dimensional output from the 12$^{\text{th}}$ Transformer layer of the HuBERT_BASE model are extracted for each spoken word at a 20 ms framerate. HuBERT_BASE model is pretrained on 960 hours of LibriSpeech data [2]. Fig. 2 illustrates the CAE-RNN model. Both the encoder and decoder are 4-layer Bidirectional-GRUs with hidden dimension of 256 and dropout of 0.2. The dimension of the AWE ($e$) is 128. The final hidden state of the encoder $h_T$ is fed to a fully connected layer $f_{enc}$ (with tanh activation) whose output is embedding $e$. Then, this embedding $e$ is replicated $T'$ times to match the target sequence length and is fed as input to the decoder [21]. Each output state of the decoder is transformed with a fully connected layer $f_{dec}$ (without activation), and its outputs constitute the final predicted sequence.

### 3.4. ASWE Implementation Details

First, AWEs are computed for the entire dataset. All the AWEs are converted into unit-length vectors $\frac{e}{||e||}$. These computed AWEs are used as input features to train the decomposition model for ASWEs as shown in Fig. 3. Each FNN ($F_1, F_2, F_3$) has 3 hidden layers with hidden units of 512, 512 and 128, respectively and dropout of 0.2. Each FNN has 152 output classes (sub-word units) as mentioned in section 3.1. However, not all the sub-words are present in the output class of the three parallel networks due to their occurrence at different position in a word. Total of 136, 152, and 141 different sub-words are present in the data at positions 1, 2, and 3. Only those sub-words will be reflected in the output class of the three networks during training. For each FNN, the output of the last hidden layer (128- dimensional) is taken as sub-word embedding representation ($e_1, e_2, e_3$) as shown in Eq. 3 and all embeddings are normalized to unit-length vectors (including reconstructed $e'$). ReLU activation is used for all layers except the last hidden layer and projection layer, where tanh activation is used to maintain the consistency between the derived AWEs and ASWEs.

## 4. EXPERIMENTS

### 4.1. Computing AWEs and Translational ASWEs

CAE-RNN model is trained with a batch size of 256 and Adam optimizer with a learning rate of 0.0001. The model is trained for 100 epochs in the case of MFCC features and

[2] https://github.com/pytorch/fairseq

20 epochs in the case of HuBERT features. In each case, the final model weights are picked based on the best AP precision value for the acoustic word discrimination task on the validation set. AWEs are extracted for the test and validation set with $ENC$ as shown in Fig. 2 and Eq. 1. The decomposition model is trained with a batch size of 256 and Adam optimizer with a learning rate of 0.001. The model is trained for 100 epochs and the final model is picked based on the best validation loss. MFCC and HuBERT-based ASWEs are extracted for the test set for evaluation.

### 4.2. AWE and ASWE Evaluation

First, to evaluate the quality of AWEs, word discrimination task is performed. The total number of generated word pairs is roughly 47 million, calculated as $\binom{M}{2} = \frac{M(M-1)}{2}$, for both the test and validation sets, each containing 9778 words. AP for the word discrimination task is calculated using these acoustic word pairs [30] for both the test and validation sets.

Then, the evaluation of ASWEs is performed, which is described as follows:

1. AP-SD: For the sub-word discrimination task, approximately 47 million sub-word pairs are generated for each position from the test set. Then, the AP-SD for each position $(p_1, p_2, p_3)$ is calculated for the test set using these sub-word pairs as described in Sec. 3.2.

2. AP-RW: For AP-RW, all words in the test set are reconstructed by translating ASWEs in the embedding space as described in Eq. 6. The spoken word pairs are generated from reconstructed AWEs ($e'$) of "spoken words" and the original AWEs ($e$) for comparison ($9778 \times 9778 \approx 95$ million). AP-RW is calculated as described in Sec. 3.2.

3. AP-CW: AWEs of written words are constructed. This is referred to as **'Word2AWE'**. Three position-dependent acoustic sub-word embedding dictionaries are derived from the three parallel networks of decomposition model $F_1, F_2$, and $F_3$. Each dictionary has sub-words as keys and the mean of all instances of their ASWEs in the training set as corresponding values. This produces three distinct sub-word dictionaries, one for each position in the word. Using these acoustic sub-word embedding dictionaries, AWEs are constructed for all the unique "written words" (3360) in the test set. First the words are tokenised and then ASWEs are obtained for those tokens from the learned dictionaries. Then translating ASWEs in embedding space would construct the AWEs as shown in Eq. 6. From 3360 constructed AWEs and a test set of 9778 original spoken word instances, $3360 \times 9778 \approx 32$ million word pairs are generated for comparison. AP-CW is calculated for these pairs as mentioned in Sec. 3.2.

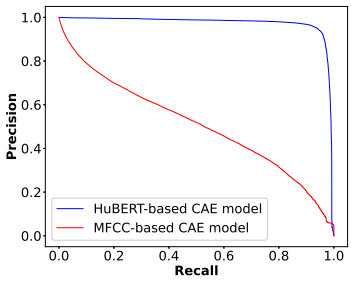


**Fig. 4**. Precision-Recall curve for the test set using a CAE-RNN model trained with HuBERT and MFCC features.

**Table 2**. Average precision for the word discrimination task on the test and validation sets for AWEs with HuBERT and MFCC as input features to CAE-RNN model.

| Input Features | Model | AP (test set) | AP (val set) |
|---|---|---|---|
| HuBERT (768-dim) | CAE-RNN | 0.97 | 0.97 |
| MFCCs with delta (60-dim) | CAE-RNN | 0.50 | 0.49 |

## 5. RESULTS AND DISCUSSION

Table 2 shows the results of the CAE-RNN model for AP on the acoustic word discrimination task. Fig. 4 shows the precision-recall curve for both the MFCC and HuBERT-based CAE-RNN models. The AP is the area under the curve, which is approximately equal to 0.97 for HuBERT-based CAE-RNN model, almost two times better than MFCC-based CAE-RNN. Fig. 5 shows the t-SNE plot of all spoken instances of the six most frequent words from the test set, derived from HuBERT-based CAE-RNN model. It is quantitatively and qualitatively evident that the derived AWEs are of high quality and well-separated in AWE space.

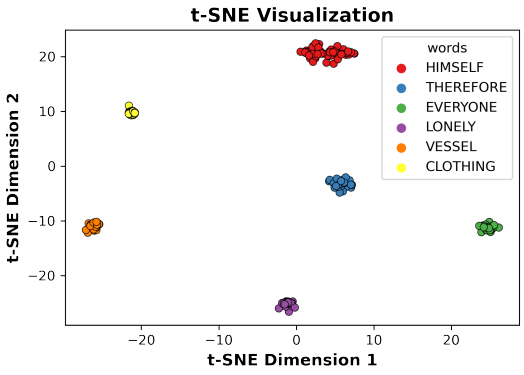


**Fig. 5**. t-SNE visualisation of AWEs of all spoken instances of the six most frequent words from the test set, derived from HuBERT-based CAE-RNN model.

**Table 3**. Performance of derived ASWEs from AWEs on the test set with various decomposition model setups.

| Loss, Features | Proj. layer | AP-SD @p1 | AP-SD @p2 | AP-SD @p3 | AP-RW | AP-CW |
|---|---|---|---|---|---|---|
| $L_a$, HuBERT | ✗ | 0.95 | 0.89 | 0.95 | 0.00 | 0.00 |
| $L_b$, HuBERT | ✗ | 0.93 | 0.87 | 0.91 | 0.95 | 0.64 |
| $L_b$, HuBERT | ✓ | 0.94 | 0.89 | 0.94 | **0.96** | **0.72** |
| $L_a$, MFCC | ✗ | 0.36 | 0.22 | 0.46 | 0.00 | 0.00 |
| $L_b$, MFCC | ✗ | 0.37 | 0.22 | 0.47 | 0.27 | 0.05 |
| $L_b$, MFCC | ✓ | 0.36 | 0.22 | 0.45 | 0.46 | 0.13 |

The results for the ASWEs are shown in Table 3 with all model variants. Table 3 shows that all the HuBERT-based decomposition model configurations are capable of producing high quality ASWEs for each position, as high values of AP-SD (up to 0.96) are obtained for them when compared with MFCC-based models. Also, adding similarity loss ($L_b$) helps to achieve a better reconstruction and translational properties as high values of AP-RW (0.95, 0.96 for HuBERT features) and AP-CW (0.64, 0.72 for HuBERT features) are obtained. Fig. 6 shows the t-SNE plot of all instances of the five most frequent sub-words from the test set, derived from the best decomposition model (AP-CW=0.72 and AP-RW=0.96). It is qualitatively evident that the derived ASWEs are of high quality and well separated in the ASWE space.

### 5.1. Qualitative Analysis of Word2AWE

AWEs of 4 random "written words" from the test set (and not seen during training) are constructed using the three sub-word dictionaries learned from the model (Sec. 4.2) with the best AP-CW (0.72), worst AP-CW (0), and with 2nd & 3rd embedding position swapped for the best model configuration (i.e. AP-CW=0.72). For these constructed words, their Top-3 nearest neighbours (Top-3 NN) from the test set are computed and listed in Table 4. Top-3 NN words for high AP-CW (0.72) are acoustically close to constructed words when compared to the low value of AP-CW (0) where Top-3 NN words are not similar to constructed words. Additionally, for the best model configuration, swapping the position of 2nd and 3rd sub-word embeddings has reduced the performance (AP-CW) from 0.72 to 0.05, and the nearest neighbours do not match with the constructed AWEs as shown in Table 4. This proves that derived ASWEs are position-dependent, which is a desired property to construct the word, as interchanging the position of the sub-words will change how the word sounds. For example, the word "aiming" can be constructed with $e_{\mathbf{a}_{p_1}} + e_{\mathbf{im}_{p_2}} + e_{\mathbf{ing}_{p_3}}$ but not with $e_{\mathbf{a}_{p_1}} + e_{\mathbf{ing}_{p_2}} + e_{\mathbf{im}_{p_3}}$, where $p_i$ represents the position of sub-words.

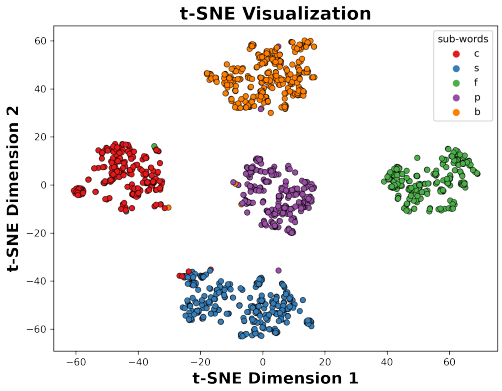


**Fig. 6**. t-SNE visualisation of ASWEs of all the instances of the five most frequent sub-words from the test set, derived from the best performing decomposition model (AP-CW=0.72).

**Table 4**. Word2AWE - AWEs of 4 random "written words" from the test set (& unseen during training) are constructed with HuBERT-based ASWEs. For these words, their Top-3 NN from the test set are shown for various model setups (multiple instances of a word could be present in the test set).

| **Words** | **vised** | **caster** | **mendes** | **indices** |
|---|---|---|---|---|
| **BPE tokens** | {*v, is, ed*} | {*c, as, ter*} | {*m, end, es*} | {*ind, ic, es*} |
| **3 Nearest neighbour (AP-CW=0.72)** | vised, vase, vested | caster, casper, cached | mendes, mades, mesas | indices, indies, indies |
| **3 Nearest neighbour (AP-CW=0.00)** | alec, snap, alack | gusto, gusto, caster | paced, bass, abasing | during, during, during |
| **3 Nearest neighbour (AP-CW=0.05)** | vase, vase, vase | covers, covers, covers | maston, amended, amended | mystic, mystic, insect |

## 6. CONCLUSION AND FUTURE WORK

Translational ASWEs are successfully extracted from high quality AWEs derived from HuBERT representations, without using any time boundaries of the spoken sub-words. Three metrics (AP-SD, AP-RW, and AP-CW) are proposed to evaluate ASWEs for the task of sub-word discrimination, reconstruction of AWEs, and construction of AWEs of "written words" by translating ASWEs in embedding space. Current work is restricted to three sub-words only (as a proof of concept). Future directions of this work include combining the AWE and ASWE extraction modules into a single framework for variable number of sub-words. These developments in the AWE and ASWE spaces are expected to help in advancing the field of query-by-example search [31, 32, 33], zero-resource speech processing [10], and learning audio-text agreement using AWEs/ASWEs for open-vocabulary keyword spotting systems [20] as a potential application of Word2AWE.

## 7. REFERENCES


[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, Red Hook, NY, USA, 2013, NIPS'13, p. 3111–3119, Curran Associates Inc.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[3] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *INTERSPEECH*, 2019.

[4] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[5] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[7] Leonardo Badino, Claudia Canevari, Luciano Fadiga, and Giorgio Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7634–7638.

[8] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux, "Unsupervised learning of acoustic sub-word units," in *ACL*, 2008.

[9] Aren Jansen and Benjamin Van Durme, "Efficient spoken term discovery using randomized algorithms," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2011, pp. 401–406.

[10] Maarten Versteegh, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux, "The zero resource speech challenge 2015: Proposed approaches and results," *Procedia Computer Science*, vol. 81, pp. 67–72, 12 2016.

[11] Keith Levin, Katharine Henry, Aren Jansen, and Karen Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 410–415.

[12] Herman Kamper, Yevgen Matusevych, and Sharon Goldwater, "Improved acoustic word embeddings for zero-resource languages using multilingual transfer," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 1107–1118, jan 2021.

[13] Yevgen Matusevych, Herman Kamper, and Sharon Goldwater, "Analyzing autoencoder-based acoustic word embeddings," *CoRR*, vol. abs/2004.01647, 2020.

[14] Shane Settle, Kartik Audhkhasi, Karen Livescu, and Michael Picheny, "Acoustically grounded word embeddings for improved acoustics-to-word speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5641–5645.

[15] Ramon Sanabria, Hao Tang, and Sharon Goldwater, "Analyzing acoustic word embeddings from pre-trained self-supervised speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[16] Badr Abdullah and Dietrich Klakow, "Analyzing the representational geometry of acoustic word embeddings," in *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022, pp. 178–191, Association for Computational Linguistics.

[17] Badr M. Abdullah, Marius Mosbach, Iuliia Zaitova, Bernd Möbius, and Dietrich Klakow, "Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study," in *Proc. Interspeech 2021*, 2021, pp. 4194–4198.

[18] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, Red Hook, NY, USA, 2013, NIPS'13, p. 2787–2795, Curran Associates Inc.

[19] Philip Gage, "A new algorithm for data compression," *C Users J.*, vol. 12, no. 2, pp. 23–38, feb 1994.

[20] Hyeon-Kyeong Shin, Hyewon Han, DoYeon Kim, Soo-Whan Chung, and Hong-Goo Kang, "Learning audio-text agreement for open-vocabulary keyword spotting," in *Interspeech*, 2022.

[21] Herman Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6535–3539, 2019.

[22] S Davis and P Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[23] Wanjia He, Weiran Wang, and Karen Livescu, "Multi-view recurrent neural acoustic word embeddings," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[24] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[25] Christiaan Jacobs, Yevgen Matusevych, and Herman Kamper, "Acoustic word embeddings for zero-resource languages using self-supervised contrastive learning and multilingual adaptation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 919–926.

[26] Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5818–5822.

[27] Herman Kamper, Weiran Wang, and Karen Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4950–4954, 2016.

[28] Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee, "Audio Word2Vec: Unsupervised Learning of Audio Segment Representations Using Sequence-to-Sequence Autoencoder," in *Proc. Interspeech 2016*, 2016, pp. 765–769.

[29] Shane Settle and Karen Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 503–510, 2016.

[30] Michael A. Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. Interspeech 2011*, 2011, pp. 821–824.

[31] Shane Settle, Keith Levin, Herman Kamper, and Karen Livescu, "Query-by-example search with discriminative neural acoustic word embeddings," in *INTERSPEECH*, 2017.

[32] Yougen Yuan, Cheung-Chi Leung, Lei Xie, Hongjie Chen, Bin Ma, and Haizhou Li, "Learning Acoustic Word Embeddings with Temporal Context for Query-by-Example Speech Search," in *Proc. Interspeech 2018*, 2018, pp. 97–101.

[33] Yushi Hu, Shane Settle, and Karen Livescu, "Acoustic span embeddings for multilingual query-by-example search," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 935–942.