IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 30, 2019, accepted January 13, 2020, date of publication January 28, 2020, date of current version February 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969983

A Synchronized Word Representation Method With Dual Perceptual Information

WENHAO ZHU, XIAPING XU, KE YAN, SHUANG LIU, AND XIAOYA YIN
Shanghai University, Shanghai 201900, China

Corresponding author: Ke Yan (664014189@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572434 and Grant 91630206, in part by the Shanghai Science and Technology Committee under Grant 19DZ2204800, and in part by the National Key Research and Development Program of China under Grant 2017YFB0701501.

ABSTRACT The information used for human natural language comprehension is usually perceptual information, such as text, sounds, and images. In recent years, language models that learn semantics from single perceptual information sources (text) have gradually developed into multimodal language models that learn semantics from multiple perceptual information sources. Sound is perceptual information other than text that has been proven effective by many related works. However, there is still a need for further research on the incorporation method for perceptual information. Thus, this paper proposes a language model that synchronously trains dual perceptual information to enhance word representation. The representation is trained in a synchronized way that adopts an attention model to utilize both text and phonetic perceptual information in unsupervised learning tasks. On basis of that, these dual perceptual information is processed simultaneously, and that is similar with the cognitive process of human language understanding. The experiment results show that our approach achieve superior results in text classification and word similarity tasks with four languages of data set.

INDEX TERMS Information representation, multi-layer neural network, natural language processing, unsupervised learning.

I. INTRODUCTION

Word representation is a method that allows a computer to understand human language by quantifying word semantics. This method is the basis of natural language processing (NLP) and has received substantial attention from academic circles.

The main way to obtain word representation is to learn the relationship between words from textual contexts [1]–[4]. However, such a distribution model only learns semantics from text, while language comprehension is more precise and includes both text and phonetics [5]–[8]. Researchers believe that when humans use language to express semantics, they will form corresponding sounds in their minds, which can help humans enhance the expression of semantics. This observation led to the development of multimodal language models that incorporate textual and sound information. A range of evaluations have shown that such models are better equipped to learn semantic word representations than text-based models [9], [10].

There are already some researches on incorporate sound in word representation. Some of them use sound as perceptual information selected sound come from natural, such as bark, running water, thunder and so on. [9], [11]. However, the representation of such sounds is limited and it cannot be used to represent abstract nouns, such as ''love'', thus greatly reducing multimodal vocabulary. The sound that can express information is phonetic, which is spoken voice. So, this paper proposes to use phonetics as the source of sound perception information.

As a general fact, the phonetic context and the text context can't be regarded as duplicated. They are a complementary relationship that provides a richer semantic for each other. For example, in the case of word sense disambiguation, ''minute'' has two meanings. when the pronunciation of ''minute'' is [ˈmɪnɪt], it indicates a time unit, and when it is pronounced [maɪˈnju:t], it means tiny. For words with similar pronunciations and different meanings, text can provide the model

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

with richer semantics (e.g., "ship" and "sheep"). Their differences in writing can help us distinguish the different meanings of the two words.

In the early stage of language learning, humans cannot express semantics in the form of textual words, but semantics are instead expressed and communicated by phonetics. The phonetic context can help people learn the language quickly. In the later stage, when humans receive both phonetic and textual information, the understanding of language is the most accurate. Nowadays, many related works usually train the two types of perceptual information separately and then incorporate the information through concatenation [12], [13]. This paper believes that the two types of perceptual information are jointly trained and incorporated during the training process, which can ensure that the language model can produce word representations with sufficient semantics.

To simulate the process of language learning in humans, this paper proposes a two-stage language model called DPWR that synchronously trains dual perceptual information and produces word representation with sufficient semantics. The first stage involves the learning of the phonetic structure. In the second stage, the attention mechanism is introduced to extract semantics from the phonetic structure of the first stage, which is incorporated with the text representation by synchronous training in the language model task. The DPWR model is compared with text-only language models and image-based multimodal language models, which reveals that the DPWR word representations are able to incorporate phonetic perception information and are superior to the word representations of other models. Moreover, compared with the simple incorporation of previous multimodal models, the incorporation method that synchronously trains dual perception information significantly improves the quality of word representation.

II. RELATED WORKS

A. PERCEPTUAL GROUNDING

There is considerable evidence from behavioral experiments and neuroimaging studies that the perceptual connection of language plays an important role in natural language processing [14]. Therefore, in recent years, a perceptually based distribution model has emerged in which language representations are learned from both textual and perceptual input. An important issue in developing such models is the source of the perceptual information.

One method for obtaining perceptual representations is to rely on direct human semantic knowledge in the form of empirically derived semantic feature production norms [15], [16], which have been used successfully in a range of multimodal models [17]–[20]. However, manual annotated norms have limited coverage and high cost. An alternative method that overcomes these restrictions is the use of raw data as the source of perceptual information. Raw data, for instance, in the form of images or sounds, are inexpensive, plentiful, and easy to obtain and provide much better coverage [21], [22].

B. MULTIMODAL MODELS

Another important issue in addition to the type of method used to capture perceptual information concerns how the two modalities (perceptual and textual) are incorporated. The multimodal representation model, which is used incorporate the two modalities, has received increasing attention. Existing integration models can generally be classified into two types.

1) JOINT TRAINING MODELS

Joint training models are models that build multimodal representations with raw inputs of both textual and perceptual resources, such as text containing images [23] or images described with text [24]. For example, Hill and Korhonen [19] proposed a corpus incorporation method that inserts the perceptual features of words into the training corpus, which is then used to train the skip-gram model [2]. These belong to early stage incorporation method, that is expanding input datas to introduce additional perceptual information. The disadvantage of this way is that not every word has corresponding perceptual information. Lazaridou et al. [25] proposed the MMSkip model, which is a medium-stage incorporation method. In their model, a convolutional neural network (CNN) is used to obtain visual features, and then, the distance between the language vector and the visual vector is minimized by means of a max-margin objective function, thereby incorporating visual information into the process of learning a language representation. The language vector and the visual vector used in this function are obtained with the skip-gram model. The works discussed above are simple extensions of the skip-gram model; they do not propose a new word embedding method. By contrast, Hu et al. [26] proposed a novel model based on a multimodal transformer architecture. Their model naturally fuses different modalities homogeneously by embedding them into a common semantic space, where self-attention is applied to model the inter- and intramodality contexts. However, this model is suitable only for visual question answering (VQA) tasks involving text.

The joint training method implicitly incorporates perceptual information into word representations while learning multimodal representation. However, these methods make use of a raw text corpus in which the words associated with the perceptual information account for a small fraction of the total words. This approach weakens the effect of introducing perceptual information, resulting in limited improvement in language representation. Therefore, some researchers have employed separate training models.

2) SEPARATE TRAINING MODELS

Separate training models are models that independently learn text representations and perceptual representations and incorporate them afterwards.

The simplest approach to incorporation following learning is to incorporate text representations and perceptual representations by concatenating them. This approach has been proven effective in learning multimodal models [9], [12], [13]. Researchers have employed transformation and dimension reduction on the concatenation results, such as singular value decomposition (SVD) [9]and canonical correlation analysis (CCA) [12]. In addition, Silberer *et al.* [27] and Silberer *et al.* [28] used superimposed multi autoencoders to learn multimodal word representations by embedding text and perceptual information inputs into a common space. However, the above methods can only generate multimodal representations of those words that contain perceptual information.

An alternative approach is to infer one modality by means of the other. Hill *et al.* [12] utilized the ridge regression method to learn a mapping matrix from the textual modality to the visual modality, and Collell *et al.* [13] employed a feed-forward neural network to learn the mapping relation between textual vectors and visual vectors. In this approach, applying the mapping function on text representations, the predicted visual vectors are obtained for all words. Then, word representations are obtained by concatenating textual and predicted visual vectors. Researchers have found that irrelevant visual information is discarded in the process of associating text to vision, which makes the predicted visual vectors outperform the original visual vectors on various semantic similarity experiments.

C. AUDITORY REPRESENTATIONS

As this work intends to show, the sources of perceptual input are not necessarily limited to visual sources. Recent work in multimodal semantics has explored the use of sound as perceptual input to learn language embeddings [11], [29]. In their work, a 'bag-of-audio-words' approach is used, in which auditory grounding is achieved by dividing sound files into frames, clustering these frames as ''audio words'' and subsequently quantizing them into representations by comparing frame descriptors with the centroids. Recently, Vijayakumar *et al.* [30] proposed an embedding scheme that learns specialized word embeddings grounded in sounds by using a variety of audio features. These techniques were found to work well for modeling human similarity and relatedness judgments and related experiments. Building on this approach, Kiela *et al.* [9] used deep learning models that led to auditory representations of higher quality.

However, the above work introduced sounds from the physical world. For example, a rumble is used as the perceptual information of the word ''thunder'', and the sound of the waves is used as perception information of the word ''sea''. However, abstract nouns, such as ''love'', do not have such physical sounds. In addition, there is work [31] believe that spoken voice carries some semantic information. Therefore, this paper uses spoken voice, i.e.,phonetics as the perceptual information and utilizes pretraining phonetic representations to enhance the effect of introducing perceptual information.

III. SYNCHRONOUS TRAIN DUAL PERCEPTUAL INFORMATION TO ENHANCE WORD REPRESENTATION(DPWR)

The model training process is divided into two stages: the phonetic structure extraction stage and the perceptual information incorporation stage. The first stage mainly obtains the phonetic structure feature of each word in the dictionary. In the second stage, unsupervised joint training of the phonetic and textual perceptual information is performed.

A. STAGE 1: PHONETIC REPRESENTATION ACQUISITION

The goal of the first phase is to encode spoken word signal features and obtain an initial phonetic representation, V_p.

We want to obtain a phonetic representation, V_p, at the word level. There are many approaches for segmenting utterances automatically. Automatic segmentation of spoken words has been successfully trained and reported previously [32], so the training audio corpus in the present work has been previously segmented into phonetic words. A word and its corresponding phonetics form a token. If two different phonetic sounds correspond to the same word (a polyphonic word), the word forms two tokens with two different phonetics to avoid the ambiguity arising from pronunciation.

For example, the textual word ''present'' has two pronunciations. When it is pronounced ['preznt], it means ''now''; when it is pronounced [prizent], it means ''give a speech''. If this textual word is considered a token, its two different semantics will be encoded into the same word representation. Regardless of the context, a word representation with two mixed semantic meanings cannot provide clear semantics. If this word is instead treated as two tokens, it will have two word representations:

''Present + ['preznt]'' = **a**
''Now + [prizent]'' = **b**

Then, when these word representations are used, the appropriate word representation can be chosen in accordance with the context or pronunciation of the word.

In addition, the audio corpus contains much unideographic noise, such as background noise and speaker characteristics. The ideographic component is the phonetic structure [31], which is not changed by the environment or the speaker. The objective of stage 1 is to disentangle the phonetic structure and noise. To achieve this objective, we designed the network structure shown in Fig. 1.

We denote the audio corpus as $X = \{x_i\}_{i=1}^{M}$,, which consists of M spoken words, each represented as $x_i = \{x_{i1}, x_{i2}, \ldots, x_{iT}\}$, where T is the total number of frames in the phonetic word, and x_{it} is the Mel-scale Frequency Cepstral Coefficient (MFCC) feature vector for the t^{th} frame. The MFCC approach is commonly used to obtain the phonetic features of audio [33]. In the MFCC approach, frequency bands are spaced along the Melscale.

The training process is divided into the two parts described below.

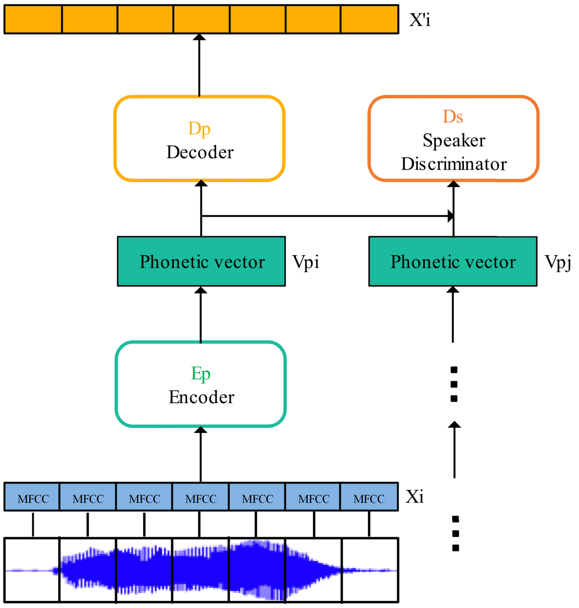


FIGURE 1. The frame of phonetic representation acquisition.

1) ENCODER AND DECODER

As shown in the left of Fig. 1, a sequence of acoustic features $x_i = \{x_{i1}, x_{i2}, \ldots, x_{iT}\}$ is entered into a phonetic encoder E_p to obtain a phonetic vector V_p. Then, the phonetic vector V_p is used by the decoder D_p to reconstruct the acoustic feature x'. This phonetic vector V_p will be used in the next stage for the phonetic representations. The encoder E_p and the decoder D_p are jointly learned by minimizing the reconstruction loss as follows:

$$L_p = \sum_i ||x_i - D_p(E_p(x_i))||_2^2 \tag{1}$$

2) PHONETIC REPRESENTATION WITH NOISE DISENTANGLED

As shown in the right of Fig. 1, a speaker discriminator D_s takes two phonetic vectors V_{pi} and V_{pj} as input and attempts to determine whether the two vectors came from the same speaker. The learning target of the phonetic encoder E_p is to "fool" the speaker discriminator D_s, preventing it from correctly discriminating the speaker identity. In this way, only the phonetic structure information is learned by E_p. The speaker discriminator D_s learns to maximize L_s in 2, while the phonetic encoder E_p learns to minimize L_s

$$\begin{aligned} L_s &= \sum_{s_i = s_j} D_s(V_{pi}, V_{pj}) - \sum_{s_i \neq s_j} D_s(V_{pi}, V_{pj}) \\ V_{pn} &= E_p(x_n) \end{aligned} \tag{2}$$

where the phonetics pi uttered by speaker s_i, $D_s\,(\cdot,\cdot)$ is a real number. [31] proved that the adversarial training framework like this can effectively remove noise.

B. STAGE 2: SYNCHRONOUSLY TRAINING OF PHONETIC REPRESENTATION AND TEXT REPRESENTATION

The unsupervised learning task of the joint training framework is similar to filling in blanks, that is, inferring the target word from the context surrounding the target word. The training framework is shown in Fig. 2. We use one sentence of window sizes as input for the stage 2. Every time input, the window moves one unit to the right. The phonetic vectors V_p obtained in stage 1 and the text vectors V_t initialized by word2vec [2] be regard as two kinds of word representations in the sentence. The input format is matrix X_{m*n}, and each row in the matrix is the n-dimension vector representation of a word. m is the context window size.

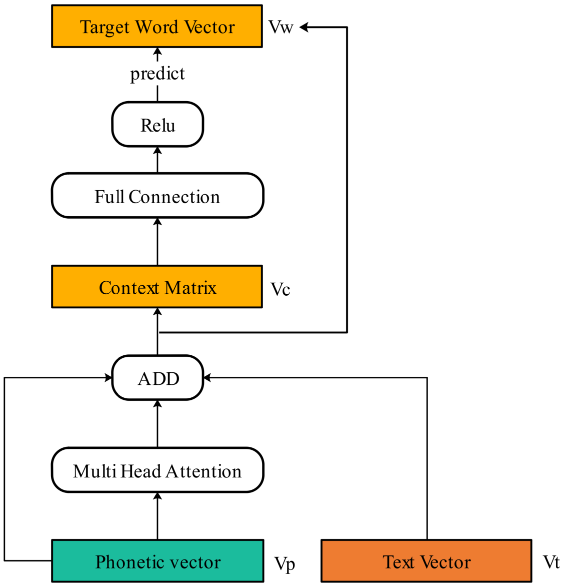


FIGURE 2. The frame of joint training.

In stage 1, only the spoken word signal feature is encoded into the phonetic vector V_p. Therefore, we need to embed the semantics in V_p in stage 2. Compared with that for textual semantics, the extraction method for phonetic semantics is more complicated. The semantics of two types of words require strengthening in the language model based on their phonetics. One is polyphonic words (such as "present" and "minute"). These words have two distinct phonetic representations but correspond to the same text representation. The other is homophones (such as "see" and "sea" or "son" and "sun"). These words have different text representations but the same phonetic representations. Therefore, we use a multihead attention model to embed the semantics into Vp. In addition, in this stage, we use residual networks to incorporate the attention matrix, the phonetic vector matrix and the text vector matrix as word representation matrices.

The n^{th} row is taken from the word representation matrix as the target word vector V_w, the other rows constitute the context matrix as input of the fully connected layer, and the output is the context representation vector V_c, which is

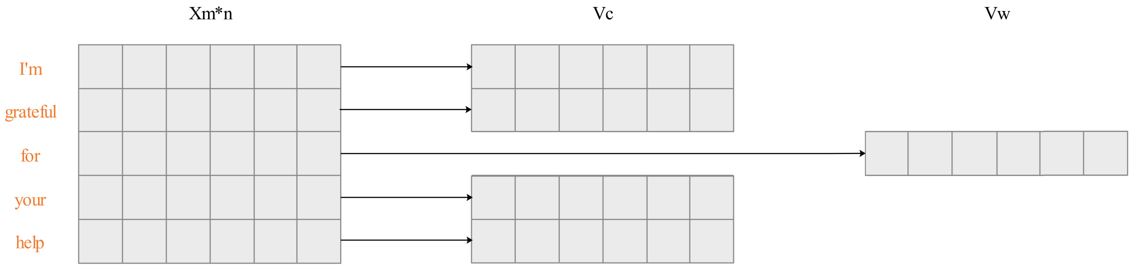


FIGURE 3. Structure of X_{m*n}, V_c and V_w.

used as the prediction target word V_w.The relationship among X_{m*n}, V_c and V_w is shown in Fig. 3

According to that two different words having similar context should have similar semantics. We extend the idea of the model and obtain the objective function for stage 2 as follows:

$$\sum_{t=1}^{T}[\sum_{c\in C_w} log(1+exp(-s(w,c))) + \sum_{n\in N_{w,c}} log(1+exp(s(w,n)))] \quad (3)$$

where C_w represents words that in the context window, and $N_{w,c}$ represents a set of negative examples sampled in the dictionary using the negative sampling method [2]. The key principle of the negative sampling method is to replace the expensive denominator with a collection of context words ‘‘negatively’’ sampled based on a distribution, which can be set as the word unigram distribution U. In practice, to overcome the data sparseness issue, we raise U to the $3/4th$ power following [2].

Extract the semantics over the phonetic representation

Reference [34] found that multihead attention allows a model to jointly attend to information from different representations subspaces at different positions. Thus, this paper utilizes a multihead self-attention model to extract phonetic contextual semantics to the phonetic representations of the current word. First, the query matrix W^Q, the key matrix W^K and the value matrix W^V of the i^{th} attention head are randomly initialized. Then, the phonetic vector matrix X^p and the parameters (W_i^Q, W_i^K, W_i^V) are input into the attention layer to obtain the attention matrix of the i^{th} attention head.

$$Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \quad (4)$$

where $Q_i = X_p * W_i^Q$; $K_i = X_p * W_i^K$; $V_i = X_p * W_i^V$; and the parameter matrices $W_i^Q \in R^{d_{model}*d_k}$, $W_i^K \in R^{d_{model}*d_k}$, and $W_i^V \in R^{d_{model}*d_v}$. d is the dimension of queries and keys. Mikolov *et al.* [2] proposed that for large values of d_k, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients. To counteract this effect, we scale the dot products by $\sqrt{d_k}$.

Finally, we linearly project the queries, keys and values h times with different, learned linear projections to d_k, d_k and d_v dimensions, respectively. On each of these projected versions of queries, keys and values, we perform the attention function in parallel, yielding d_v-dimensional output values. This output attention matrix, the phonetic representations matrix X_p and the text representations matrix X_t are concatenated and once again projected, resulting in the final values, as depicted in the lower half of Fig. 2.

$$\begin{aligned} Multihead(Q, K, V) &= Concat(X_p, X_t, head_1, \ldots, head_h)W^o \\ head_i &= Attention(Q_i, K_i, V_i) \\ W^o &\in R^{(h+2)*d_v*d_{model}} \end{aligned} \quad (5)$$

IV. EXPERIMENTS

In this section, we test the effectiveness of the DPWR model in generating high-quality word representations.

A. DPWR MODEL SETTINGS AND DATASETS

Our experiments cover four languages: English, Spanish, German, and French.

The training textual corpus we use is comprised of Wikipedia datasets in English,[1] Spanish,[2] German[3] and French.[4] We have dictionary sizes of 0.34, 0.38, 0.16, and 1.40 million words for English, Spanish, German and French, respectively.

The English audio corpus was LibriSpeech [35], which is a corpus of read speech in English derived from audiobooks. This corpus contains 1000 hours of speech sampled at 16 kHz uttered by 2484 speakers. The audio corpus for the other languages was the GlobalPhone [36], a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in 20 languages. This corpus contains 400 hours of transcribed audio data from more than 2000 native speakers. We extracted 39-dimension MFCCs as the acoustic features.

[1] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

[2] https://dumps.wikimedia.org/eswiki/latest/eswiki-latest-pages-articles.xml.bz2

[3] https://dumps.wikimedia.org/dewiki/latest/dewiki-latest-pages-articles.xml.bz2

[4] https://dumps.wikimedia.org/frwiki/latest/frwiki-latest-pages-articles.xml.bz2

In stage 1, the phonetic encoder E_p and decoder D_p both are 2-layer GRUs with hidden layer size 128 and 256, respectively. The speaker discriminator D_s is a fully-connected feed-forward network with 2 hidden layers with size 128. The size of representation vectors is set to 128.

In stage 2, the number of self-attention heads is set as 4, and the parameters of each attention layer are as follows: $W_i^Q \in R^{512\times128}$, $W_i^K \in R^{512\times128}$, $W_i^V \in R^{512\times128}$ and $W^o \in R^{768\times512}$. The fully-connected feed-forward network has 2 hidden layers with size 258. The size of representation vectors is set to 128. The context window size is 5, and the negative sampling number is 5.

B. TRAINING SCHEMES

To examine whether the pretraining of phonetic structure in stage 1 and the synchronously training in stage 2 improve performance in word representations, we design three model training schemes. The first is DPWR, which performs phase one and phase two of model training. The second is DPWR-stage2, which directly concatenates text and phonetic representations after stage 1. The third is DPWR-stage1, which does not learn phonetic structure features before synchronously training.

C. BASELINE ALGORITHMS

To assess the effectiveness of DPWR, we compare several text-only models, such as sisg [3], word2vec [2] (including the skip-gram and CBOW algorithms), and the GloVe model [37].

In addition, we compared DPWR with text+boaw and text+nae, which were proposed by Kiela *et al.* [9]. Both are auditorily grounded multimodal models. These two models achieve multimodal word representations by concatenated representations of actual sound and text. The audio data are from the online search engine Freesound[5]. text+boaw makes use of the so-called ''bag of audio words'' (BoAW) algorithm to obtain auditory-grounded representations. text+nae makes use of a neural network to obtain auditory-grounded representations.

In addition, we compared DPWR with TunedFL and TunedSL [9], which are both image-based multimodal language models that use visual information as perception information. TunedFL assigns equal weights to the textual and visual components, and TunedSL uses the Scoring Level (SL) strategy (with similar weights assigned to the two channels, and the same k values as TunedFL).

D. EVALUATION METHODS

Schnabel et al. [38] show that the word similarities task is a prominent means for evaluating word representations. The word similarities task is an intrinsic evaluation method. Intrinsic evaluation methods concentrate on measuring lexical internal patterns such as semantic and morphology information. A language model that achieves good performance at intrinsic evaluation cannot produce similar performance in extrinsic evaluation. So, in addition, we use the text classification task for extrinsic evaluation.

E. WORD SIMILARITY TASK

1) EXPERIMENTAL SETTINGS AND DATASETS

We assess the performance of DPWR in predicting the degree of semantic relatedness between two words as rated by human judges, and compared the four DPWR model training schemes with baseline models in English.

We evaluated DPWR, purely textual models, image-based multimodal models, and sound-based multimodal models on several standard word similarity datasets, such as WS-353 [39], MC-30 [40], RG-65 [41] and MEN-TR-3000 [42]. Each word pair in these datasets is associated with several human judgments on similarity and relatedness on a scale from 0 to 10 or 0 to 4.

WordSim353 is a widely used benchmark constructed by asking 13 subjects to rate a set of 353 word pairs on an 11-point meaning similarity scale and averaging their ratings (e.g., dollar/buck receives a very high average rating, professor/cucumber receives a very low one).

Among the word similarity datasets we use, MEN was developed specifically for the purpose of testing multimodal models. It consists of 3000 word pairs with [0, 1]-normalized semantic relatedness ratings provided by Amazon Mechanical Turk workers. For example, beach/sand has a MEN score of 0.96, bakery/zebra received a 0 score. Since the MEN dataset comprises a wide variety of concepts, one could argue that this dataset was appropriate for our purposes.

The models are evaluated as follows. For each pair in a data set, we compute the cosine of the model vectors representing the words in the pair and then calculate the Spearman correlation of these cosines with the human ratings of the same pairs. The higher the correlation, the better the model can score in correlation analysis, and the more accurate the word representations can be. When a word has two representations (i.e., two pronunciations), we choose the one that performed better in the word similarity experiment.

It can be observed from Table 1 that the experimental results of DPWR are superior to those of the other models. Therefore, as reported in Table 2, this paper also conducts the word similarity experiment in other languages and compares the DPWR model with the baseline models. We use WS353, MC and RG for the Spanish models, German WS353 for the German models, and French WS353 for the French models.

2) EXPERIMENTAL RESULTS

The experimental results of the word similarity tasks are reported in Table 1 and Table 2.

As shown in Table 1, the perceptual-based model has no semantic word representations that are worse than those of the purely textual models regardless of the source of perceptual information. The results show that the method of extracting semantics from perceptual information and the incorporation

[5] http://www.freesound.org.,Roma & Serra, 2013

TABLE 1. Spearman's Correlation $\rho \times 100$ for word similarity datasets(%) in english.

Evaluation	Text			Text+Sound					Text+Image	
	sisg	CBOW	skip-gram	text +boaw	text +nae	DPWR	DPWR-stage1	PATE-stage2	TunedFL	TunedSL
WS-353	72.57	67.61	71.64	71.22	72.52	**75.22**	72.15	66.13	72.05	71.12
MEN-TR-3000	72.81	58.02	65.98	68.94	67.30	**78.80**	72.78	70.45	78.29	78.05
MC-30	70.78	62.91	67.49			**75.82**	73.28	62.18		
RG-65	74.76	60.41	66.63			**78.80**	75.14	70.74		

TABLE 2. Spearman's Correlation $\rho \times 100$ for word similarity datasets(%) in other languages.

Lang.	Evaluation	sisg	CBOW	skip-gram	DPWR
Spanish	WS-353	58.27	32.92	50.15	**59.32**
	MC-30	77.54	60.12	72.22	**79.57**
	RG-65	82.97	48.47	74.64	**83.94**
German	MWS353-German	52.54	46.35	48.59	**55.72**
	WS353-German-rel	52.10	40.77	47.2	**54.67**
	WS353-German-sim	55.66	42.61	53.31	**58.13**
French	WS-353	51.34	41.46	44.77	**53.66**

method of perceptual information can affect word representations quality.

When comparing sound-based representations with image-based representations, We found that on the MEN-TR-3000 dataset, word representations that incorporate image perceptual information are of higher quality than those that simply incorporate sound perceptual information, but the DPWR has better performance. Comparison among the several models of audio-based representations reveals that the DPWR model with synchronously training perceptual information can produce the better word representations, especially on the MEN-TR-3000 dataset.

By comparing the three models DPWR, DPWR-stage1 and DPWR-stage2, it can be seen that the model training scheme has an effect, and the word representations of DPWR model has the highest quality. The quality of DPWR-stage1 is worse than that of DPWR, indicating the importance of the pre-trained phonetic structure. DPWR-stage2 is the worst of the three models, indicating that the incorporation method of multimodal language model plays a large role in determining the quality of word representations. synchronously training perceptual information during the process of language model learning is more effective than simple concatenate.

In other languages, DPWR's performance is 2% to 10% better than that of other models. The results demonstrate that incorporation method of DPWR is equally valid for other languages.

F. TEXT CLASSIFICATION TASK

In the text classification experiment, it is appropriate to select the word representations of the polyphonic word by context. We choose several models that perform well in word similarity experiments and perform text classification task on these models. We evaluate the performance of our models in two languages.

TABLE 3. Accuracy (%) of text classification in each language.

Model	Yelp review (English)	TASS 2017 (Spanish)
CBOW	80.77	71.61
Skip-Gram	79.92	70.37
Glove	80.17	73.04
sisg	90.1	73.74
DPWR	89.8	75.52

1) EXPERIMENTAL SETTINGS AND DATASETS

For English, we use Yelp reviews as a classification data set. We use the Yelp reviews from [43]. There are 1,569,264 samples of Yelp reviews with a total of 4 ratings. In this experiment, we consider the first level and second level as negative and the third level and fourth level as positive. For Spanish, we choose TASS 2017 [44] as a text classification data set. TASS 2017 is an emotional classification dataset based on Spanish Twitter that includes four categories: "P", "N", "NEU", and "NONE". We use the average of the word vectors contained in the text to represent the text. When a word has two representations (that is, two different pronunciations), we select one that is closer to the context representation as the current word representation, and the calculation is as follows:

$$f(w) = \sigma(x_u^T \theta_w), u \in context(w) \tag{6}$$

where x_u represents the sum of the word vectors of all words in context(w), and θ_w represents the word vectors of the word w.

Text classifier training is performed using the LIBLINEAR tool [45]. For the data sets that do not have separate training and test sets, we select 70% of the data as the training set and 30% as the test set. The experimental results are reported in Table 3.

TABLE 4. Nearest neighbors of "minute" using TunedFL and DPWR.

TunedFL	DPWR	
seconds, scored, half, extra, goal, substitute, quarter, coming, penalty, final	[’mɪnɪt]	half, score, extra, twice, average
	[maɪ’nju:t]	Draw, drop, quick, fewer, size

2) EXPERIMENTAL RESULTS

As can be observed from Table 3, the DPWR model achieved the superior results in each language. In English, the accuracies of sisg and DPWR reach 89%, but that of sisg is 0.05% greater than that of DPWR.In Spanish, PATE has an accuracy 1.78% higher than the second. sisg is tied for second, with an accuracy rate of 73.74%.

G. QUALITATIVE ANALYSIS

Comparing the five models that incorporate the phonetic features listed in Table 1 shows that DPWR and DPWR-stage1 outperform text+boaw and text+nae, which combine two features in a proportionally concatenated manner. While in the DPWR and DPWR-stage1 models, the semantic meaning of perceptual information is extracted and incorporated by synchronously training in neural network. This suggests that the incorporation method of DPWR is more effective than separately training and concatenation. Different incorporation methods lead to differences in the results.

Comparing the experimental results of image-based and sound-based multimodal language models, reveals that the sound-based models have no marked advantages. We analyze semantic relations of word representations, hoping to obtain different results.

We use the word "minute" as an example. In the feature space obtained by the TunedFL and DPWR models, we remove the morphological inflections on the word ’minute’, such as ’miniutes’, ’minuted’, and ’minuter’, and find the ten nearest neighbors of "minute". The nearest neighbors of a word are computed by comparing the cosine similarity between the center word and all other words in the dictionary. The results are reported in Table 4. The image-based multimodal model TunedFL indicates that some of the nearest neighbors are seemingly sports related. The TunedFL model infers words related to time based on the meaning of "minute". The nearest neighbors of the two pronunciations "minute" in the DPWR model show that the different pronunciations allows the contexts of different meanings of a word to be classified into separate groups, allowing our model to learn multiple meanings of a word.

V. CONCLUSION AND FUTURE WORK

Most of the language models that incorporate sound as perceptual information into word representations do not apply a good incorporation method, which leads to limited improvement in the quality of language model. Inspired by the human language learning process, this paper presents the language model named DPWR, which synchronously trains dual perceptual information to enhance word representation. The representation is trained in a synchronized way that adopts an attention model to utilize both text and phonetic perceptual information in unsupervised learning tasks. On basis of that, these dual perceptual information is processed simultaneously.

The above experiments suggest that the method used to perceptual information, the method used to extract perceptual information, and the source of perceptual information all affect the quality of the language model. The experimental results show that the incorporation method of joinly training is more effective than the separately training and simple concatenation. In addition, pretraining of the phonetic structure can improve the quality of the language model.

As one of the main research directions related to the development of language representations, the performance of multimodal language models depends not only on the source of perceptual information but also on the method used to incorporate that information. Such an incorporation method should not be limited to the incorporation of only two kinds of information but should also be capable of incorporating information from more than two modes. This work initially explores a method of training a multimodal language model to incorporate phonetic perceptual information. In future work, we will continue to explore incorporation methods for multimodal information and use these methods for sentence- and chapter-level representation.

REFERENCES

[1] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Intell. Res.*, vol. 37, pp. 141–188, Jul. 2018.

[2] T. Mikolov, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[4] W. Zhu, T. Yao, J. Ni, B. Wei, and Z. Lu, "Dependency-based Siamese long short-term memory network for learning sentence representations," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, Art. no. e0193919.

[5] C. A. Perfetti, "The limits of CO–occurrence: Tools and theories in language research," *Discourse Processes*, vol. 25, nos. 2–3, pp. 363–377, Jan. 1998.

[6] W. Zhu, X. Jin, J. Ni, B. Wei, and Z. Lu, "Improve word embedding using both writing and pronunciation," *PLoS ONE*, vol. 13, no. 12, Dec. 2018, Art. no. e0208785.

[7] A. M. Glenberg and P. Michael Kaschak, "Grounding language in action," *Psychonomic Bull. Rev.*, vol. 9.3, no. 2002, pp. 558–565.

[8] M. H. Bornstein, L. R. Cote, S. Maital, K. Painter, S.-Y. Park, L. Pascual, M.-G. Pecheux, J. Ruel, P. Venuti, and A. Vyt, "Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, french, hebrew, italian, korean, and american english," *Child Develop.*, vol. 75, no. 4, pp. 1115–1139, Jul. 2004.

[9] D. Kiela and S. Clark, "Learning neural audio embeddings for grounding semantics in auditory perception," *J. Artif. Intell. Res.*, vol. 60, pp. 1003–1030, Jul. 2018.

[10] P. Gurunath Shivakumar and P. Georgiou, "Confusion2Vec: Towards enriching vector space word representations with representational ambiguities," *PeerJ Comput. Sci.*, vol. 5, p. e195, Jun. 2019.

[11] A. Lopopolo and E. van Miltenburg, "Sound-based distributional models," in *Proc. 11th Int. Conf. Comput. Semantics*, Apr. 2015, pp. 70–75.

[12] F. Hill, R. Reichart, and A. Korhonen, "Multi-modal models for concrete and abstract concept meaning," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 285–296, Dec. 2014.

[13] C. Collell, T. Zhang, and M.-F. Moens, "Imagined visual representations as multimodal embeddings," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 4378–4384.

[14] L. W. Barsalou, "Grounded cognition," *Annu. Rev. Psychol.*, vol. 59, pp. 617–645, Jan. 2008.

[15] D. L. Nelson, C. L. Mcevoy, and T. A. Schreiber, "The university of South Florida free association, rhyme, and word fragment norms," *Behav. Res. Methods, Instrum., Comput.*, vol. 36, no. 3, pp. 402–407, Aug. 2004.

[16] K. Mcrae, G. S. Cree, M. S. Seidenberg, and C. Mcnorgan, "Semantic feature production norms for a large set of living and nonliving things," *Behav. Res. Methods*, vol. 37, no. 4, pp. 547–559, Nov. 2005.

[17] B. T. Johns and M. N. Jones, "Perceptual inference through global lexical similarity," *Topics Cognit. Sci.*, vol. 4, no. 1, pp. 103–120, Jan. 2012.

[18] S. Roller and S. S. I. Walde, "A multimodal LDA model integrating textual, cognitive and visual modalities," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2013, pp. 1146–1157.

[19] D. Kiela, F. Hill, A. Korhonen, and S. Clark, "Improving multi-modal representations using image dispersion: Why less is sometimes more," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, Jun. 2014, pp. 835–841.

[20] L. Bulat, D. Kiela, and S. Clark, "Vision and feature norms: Improving automatic feature norm learning through cross-modal maps," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Jun. 2016, pp. 579–588.

[21] M. Baroni, "Grounding distributional semantics in the visual world," *Language Linguistics Compass*, vol. 10, no. 1, pp. 3–13, Jan. 2016.

[22] S. Wang, J. Zhang, and C. Zong, "Learning multimodal word representation via dynamic fusion methods," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 5973–5980.

[23] Y. Feng and M. Lapata, "Visual information in semantic representation," in *Proc. Hum. Lang. Technol. Annu., Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Jun. 2010, pp. 91–99.

[24] G. Synnaeve, M. Versteegh, and E. Dupoux, "Learning words from images and speech," in *Proc. NIPS Workshop Learn. Semantics*, Montreal, QC, Canada, 2014, pp. 1–5.

[25] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," Jan. 2015, *arXiv:1501.02598*. [Online]. Available: https://arxiv.org/abs/1501.02598

[26] R. Hu, "Iterative answer prediction with pointer-augmented multimodal transformers for text VQA," *arXiv:1911.06258*. [Online]. Available: https://arxiv.org/abs/1911.06258

[27] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Jun. 2014, pp. 721–732.

[28] C. Silberer, V. Ferrari, and M. Lapata, "Visually grounded meaning representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2284–2297, Nov. 2017.

[29] D. Kiela and S. Clark, "Multi- and cross-modal semantics beyond vision: Grounding in auditory perception," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2461–2470.

[30] A. K. Vijayakumar and R. V. D. Parikh, "Sound-word2vec: Learning word representations grounded in sounds," Mar. 2017, *arXiv:1703.01720*. [Online]. Available: https://arxiv.org/abs/1703.01720

[31] Y.-C. Chen, "Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 941–948.

[32] Y.-H. Wang, H.-Y. Lee, and L.-S. Lee, "Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6269–6273.

[33] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[34] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[35] V. Panayotov, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[36] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8126–8130.

[37] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[38] T. Schnabel, "Evaluation methods for unsupervised word embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 298–307.

[39] L. Finkelstein, "Placing search in context: The concept revisited," *ACM Trans. Inf. Syst.*, vol. 20, no. 1, pp. 116–131, 2002

[40] G. A. Miller and G. W. Charles, "Contextual correlates of semantic similarity," *Language Cognit. Processes*, vol. 6, no. 1, pp. 1–28, 1991.

[41] H. Rubenstein, and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965.

[42] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artif. Intell. Res.*, vol. 49, pp. 1–47, Jan. 2014.

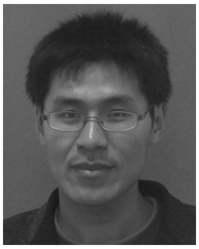
[43] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 649–657.

[44] E. Martínez-Cámara, "Overview of TASS 2017," in *Proc. TASS*, Sep. 1896, pp. 13–21.

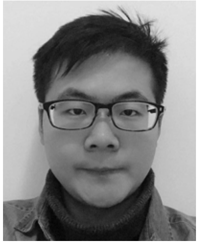
[45] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

WENHAO ZHU was born in September 1979. He received the bachelor's, master's, and Ph.D. degrees from Zhejiang University, in 2002, 2006, and 2009, respectively. From 2012 to 2013, he visited the Computer Laboratory, University of Cambridge as a Visiting Scholar. He is currently an Associate Professor with the School of Computer Engineering and Science, Shanghai University, China. His researches are in the areas of text representation, information extraction, and web data mining.

XIAPING XU is currently pursuing the master's degree with Shanghai University. Her main research fields include artificial intelligence, natural language processing, and machine learning.



KE YAN is currently pursuing the Ph.D. degree with Shanghai University. His main research fields include natural language processing and integrated circuit design.



SHUANG LIU is currently pursuing the master's degree with Shanghai University. His main research fields include artificial intelligence, natural language processing, and machine learning.

XIAOYA YIN is currently pursuing the master's degree with Shanghai University. Her main research fields include artificial intelligence, natural language processing, and machine learning.