# PHONOLOGICALLY-ENHANCED CHARACTER EMBEDINGS BY TOBIAS RAMOSER

**Anton Serjogin**
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

12.06.2017

Applications of machine language processing play an important role in our everyday lives. Most of the approaches are word-based, however, this work represents a different approach - a phonological vector representation of letters.

The goal is to create different vector representation of letters with morphological features and then compare random vectors with the transciptions in SAMPA. Which is why the following theoretical background is needed: phonetics and phonology, Word2Vec, SAMPA-Alphabet. Phonology descripes the semantics of sound within a language (in this case German) and the phonemes must differ in a phonetic property. Articulation and voicing can be divided into 3 groups: manner of articulation ( plosive, fricative, nasal, lateral, vibrant, approximant), area of articulation (bilabial, labiodental, alveolar, dental, velar, glottal), voicing (voiced, unvoiced).

Word2Vec is a program that creates vectors from words, where training data is used as an input and word vectors and distance to a word are outputted. The processing is done through a neural network, consisting of architecture and a learning algorithm, where similar words have similar vectors. One of the models is the Continuous Bag-of-Words model is a two-layered neural network that predicts a certain word ahead, depending on the context. Another model is the Skip-Gram model, a two-layered neural network as well as the Continuous Bag-of-Words, however, the difference is that it predicts the context from the given word. Following learning algorithms are used in order to reduce the training time: hierarchical softmax, negative sampling, downsampling of frequent words. SAMPA-Alphabet is ASCII-based and machine-readable, with which the pronounciation of sounds is represented.

Following the experimentation part, it is divided overall into four experiments:

- Implementation of vector creation + random vectors
- Word2Vec vectors
- Transkription into SAMPA

First of all, the initialization of a vector is made (0, 0, 0 ,0 ,0), where phonological features and sub-categories are prior defined. Then the calculation of phoneme vectors takes place and the properties are applied to the corresponding vector. The letter vector is calculated by means of the average of all the corresponding phoneme vectors. One-hot vectors contain similar features and sub-categories as char-vectors, but the classification is binary and the vector size ”grows” up to 22 dimensions. Randomized vectors are random vectors generated from two types of char-vectors (15 and 100 dimensions), where the ”random”-nupy modul is used. The Word2Vec vectors used letters instead of words as training data, the phonologically similar letters are trained on each other with certain parameter setting, this is also made in two versions: 15 and 100 dimensions. From the qualitative analysis a hunder-dimension vector space has a much better performance, than the fifteen one.