Modeling exemplar production from speech tokens with acoustic targets

Frédéric "Fred" Mailhot
Dialpad, Inc.
fred.mailhot@dialpad.com

Cassandra L. Jacobs
Department of Linguistics
University at Buffalo
cxjacobs@buffalo.edu

1 Introduction

We present a solution to the problem of exemplar production from variable-duration tokens. Incorporating time-series alignment and clustering algorithms, our model MNEMORPHON[1] stores and outputs tokens of phonetically detailed acoustic representations of recorded speech. We show qualitatively that model outputs retain high-level phonetic characteristics, and quantitatively that they contain sufficient detail for statistical classification.

2 Time in exemplar models

Within and across speakers, distinct utterances of a given word vary widely in duration, as shown in Figure **??**; this variability is a core obstacle to modeling exemplar-based production from tokens of real speech, e.g. raw audio, or spectrograms.[2]

A central component of any exemplar-based model is a means of computing distance or similarity between exemplars (Johnson, 2007). This computation must be robust to length-wise variations; a naive application of point-wise distance computation on unaligned inputs is likely to result in uninterpretable measurements of outputs. Such an algorithm in fact already exists in the speech recognition and time series literature; we show below how to incorporate it into an exemplar production model.

2.1 Distances between unequal-length sequences

Dynamic time warping (Vintsyuk, 1968; Sakoe and Chiba, 1978, DTW) is an algorithm for computing the distance between a pair of discrete sequences of potentially differing lengths. Given sequences $X = (x_0, ..., x_m)$ and $Y = (y_0, ..., y_n)$ whose elements are embedded in a shared parametric space M equipped with a distance function $d_M(x_i, y_j)$, DTW finds the optimal alignment between X and Y via the following minimization:

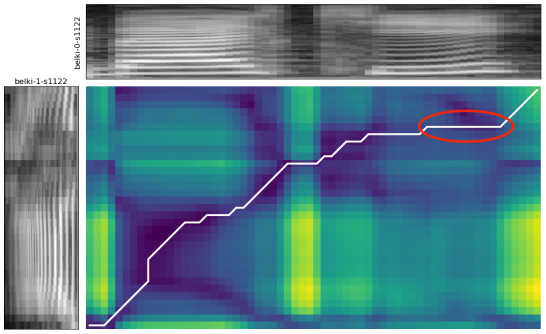

Figure 1: DTW alignment between two tokens of the Turkish word *belki* ("maybe") uttered by the same speaker. Each pixel in the grid represents the Euclidean distance between the corresponding frames of the mel spectrograms on the axes; darker pixels are closer. The white line represents the alignment path that minimizes the sum of frame-wise distances, the circled section illustrates an instance of many-to-one alignment.

$$DTW_\pi(X, Y) = \arg\min_\pi \sqrt{\sum_{(i,j)\in\pi} d_M(x_i, y_j)^2} \quad (1)$$

where π is an alignment or *warping path*, a sequence of pairs $[(i_0, j_0), ..., (i_k, j_k)]$ whose members are indices of positions in X and Y, respectively, subject to conditions on monotonicity, endpoint index alignment, and exhaustiveness. The *DTW distance* between X and Y is the sum of element-wise distances $d_M(x_i, y_j)$ over the optimal alignment.

Intuitively, warping in DTW corresponds to stretching or compressing signals along the temporal dimension. For the discrete case, this corresponds to single indices in one sequence being aligned with multiple indices in the other, as illus-

[1] Code for all experiments described here will be available at https://github.com/calicolab/mnemorphon.

[2] We focus on speech here but believe our approach applies *mutatis mutandis* to signed languages.

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 281-284.
Irvine, California, June 27-29, 2024

trated in Figure 1.[3]

3 Generalization in exemplar production

The task of exemplar production is to generate an appropriate output token, given a target category of previously stored exemplars. One trivial method is to select a memorized exemplar from the category and directly output it. On this approach the model becomes a look-up table, incapable of outputting a form it has not previously stored. A core desideratum of linguistic theories and models is the capacity to generalize beyond prior experience, a hallmark of linguistic cognition.

Pierrehumbert (2001) presents a model that implements a simple but effective method of exemplar composition for generalization in production. Given a vowel category V instantiated by points in formant space, generation of an output exemplar proceeds as follows: first, a *seed* exemplar v is randomly selected from V, next all exemplars within a fixed distance of v are collected into an exemplar *cloud*, and finally an output exemplar $\hat{v}$ is generated by computing a weighted average of the cloud. Mailhot (2010) extends this approach to fixed-length sequential tokens in an exemplar production model of vowel harmony; below we draw inspiration from these approaches and show how they can be extended to sequences of varying lengths.

3.1 Averages of unequal-length sequences

Our task is to generate an average of a set of variable-length exemplars, a problem that is known to be computationally intractable (Elias, 2006). Petitjean et al. (2011) introduce *DTW barycenter averaging* (DBA), a theoretically motivated approximation method.

Given a set of sequences $\mathcal{S}$ and an initial "best-guess" sequence $\hat{s}$ (randomly generated or sampled from $\mathcal{S}$), DBA repeatedly iterates over two phases: (i) compute $DTW_{path}(\hat{s}, s)$ for each $s \in \mathcal{S}$ and for each element of $\hat{s}$ store the set of elements from each s it was aligned with, (ii) update each element of $\hat{s}$ to be the centroid of its associated coordinates from the alignment phase.

Iterating over these phases progressively converges to a locally optimal *barycenter* $\hat{s}^*$ of $\mathcal{S}$, the sequence that minimizes the sum of squared DTW distances to all $s \in \mathcal{S}$:

$$\hat{s}^* = \arg\min_{\hat{s}} \sum_{s \in \mathcal{S}} DTW(\hat{s}, s)^2 \qquad (2)$$

DBA is guaranteed to converge as the quantity in Equation 2 stays the same or decreases at each iteration; the update either moves the barycenter's coordinates to be closer to their aligned cloud elements, or else a lower-cost DTW alignment is found.

3.2 MNEMORPHON

With the above pieces in place, we introduce our own model of exemplar production. MNEMORPHON is intended as a partial model of human phonology. It stores word-sized exemplars, encoded as mel-scaled spectrograms of recorded speech, along with a quasi-phonemic string representation of the word category. There is no representation of sub-lexical units such as segments or syllables.

MNEMORPHON uses DBA as its core production algorithm, with the possibility of similarity-weighting by inverse DTW distance. Note that DBA is not applied to each mel band individually as this would likely cause alignment issues. Instead the alignment is frame-wise, which each slice of the spectrogram considered as a single element of a multi-dimensional time series.

4 Experiments

We conduct two sets of experiments: the first investigates the quality of DBA-generated outputs and its dependence on cloud size, and the second shows that MNEMORPHON's outputs contain sufficient information to be accurately phonetically categorized.

4.1 The data

For the experiments described below our raw data set is a corpus of Turkish speech.[4] The corpus comprises microphone recordings from 120 speakers who each read 40 sentences sampled from a triphone-balanced set of 2462 Turkish sentences (16KHz sample rate, balanced across binarized gender; age 19–50 years, mean=23.9).[5] Each recorded sentence is transcribed in standard Turkish orthography as well as an ASCII-compatible phonemic orthography called *METUbet* (Özgül Salor et al.,

[3] Kirchner et al.'s (2010) PEBLS exemplar model also makes use of DTW, although its production algorithm differs significantly from MNEMORPHON's.

[4] https://catalog.ldc.upenn.edu/LDC2006S33

[5] Additional metadata for each speaker includes dates of birth and recording, places of birth and residence, and level of education.

2002). The corpus also includes word- and phone-level alignments.

Inspection revealed a subset (n=23) of the speakers in the corpus to have mismatches between audio and transcript files. These were filtered out, leaving 97 speakers (m=49, f=48) for all experiments described below.

4.2 Data processing

As mentioned, MNEMORPHON's inputs are words; these are segmented from the corpus speech files using the provided word-level alignments. Each segmented word is stored with its METUbet string representation as category label, along with speaker ID, gender marker, and a within-speaker token index. The segmented word audios are then encoded as mel-scaled spectrograms, with the following parameters:

- window length: 46ms
- hop length: 12ms
- 80 mel bands

As can be seen in Figure 1, these spectrogram parameters generate comparatively coarse *narrowband* spectrograms. Our choice of spectrogram parameters was constrained by our evaluation methodologies, discussed below.

For our second experiment we used the phone-level alignments to segment vowel tokens from our training corpus and MNEMORPHON's outputs.

4.3 Experiment 1: Averaging as an output strategy

For each word in our corpus with ten or more associated exemplars, we uniform randomly selected one token as the seed exemplar and used subsets of varying sizes of the remaining tokens as the "cloud" from which a barycenter was computed; the sampling and averaging blind to speaker identity or indexical information such as gender. Figures 2 and Figure 3 plot properties of the seed and DBA-computed outputs for various cloud sizes for a representative token.

At the global level we see that the general acoustic properties (e.g. areas of high or low energy across different frequencies and frame sequences) are retained, although DBA clearly introduces noise as cloud size increases. Inspection of individual mel bands shows that DBA is able to compute a meaningful average for temporally variable signals, reliably locating the main energy peak and troughs.

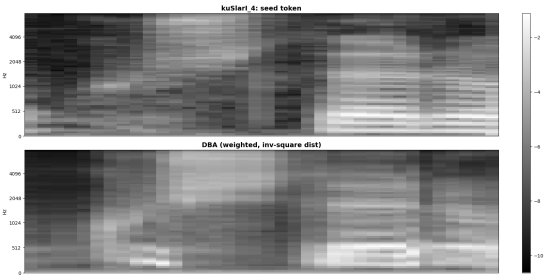

Figure 2: Seed and MNEMORPHON output spectrograms (cloud size: 20) for *belki* exemplar.

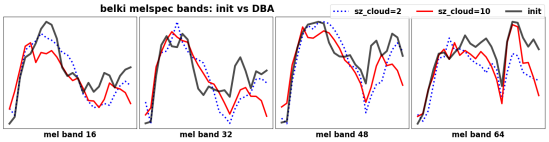

Figure 3: Mel bands 16, 32, 48, 64 of seed and DBA-generated spectrograms for *belki*.

4.4 Experiment 2: Output categorization

Here we ask whether slices of DBA-generated outputs retain sufficient phonetic detail to be statistically classified according to a standard phonetic feature, vowel frontness/backness.[6]

Although MNEMORPHON has no representation of sub-lexical units, they are useful in the context of this extrinsic analysis. For this experiment, we extracted all vowels from the audio corpus using the included alignments and labeled them according to the METUbet vowel symbol they were transcribed as. We converted them directly to mel spectrograms as above, resulting in a total of 82360 samples, which were randomly shuffled and divided via stratified split into train, development, and test sets representing 80, 10, and 10 percent of the corpus samples.

Our classifier is a convolution neural network. They are known to perform well on spectrograms and in fact form the backbone of many contemporary speech recognition systems (Gulati et al., 2020). Our network has 4 layers of 2-d convolutions (5x5 in the first layer and 3x3 for subsequent layers), a max-pooling layer, and a final fully-connected layer projecting to a binary output (representing [± back]). Kernel sizes, learning rate and batch size were tuned on the development split; the final training run was for 25 epochs.[7]

[6] We examine this feature in particular in the context of in-progress work assessing MNEMORPHON's ability to model vowel harmony and other morphophonological alternations.

[7] The accompanying repository has fuller details of the

4.4.1 Data augmentation

Neural networks, like other supervised learners, are sensitive to distribution shift, where features relevant to classification are differently distributed in the training and evaluation sets. This is the situation in the current experiment; our training data consists solely of "clean" spectrograms directly computed from audio while the target spectrograms are "noisy" for reasons discussed above. For this reason our initial attempts at classifying MNEMORPHON's outputs fared poorly, with performance at or near chance.

In order to mitigate the effect of this disparity we augmented our training data with DBA-generated samples; for each vowel category we added 1000 samples, each created by running distance-weighted DBA over 10 tokens uniform randomly sampled from the given category's exemplars in the training set.

4.5 Results

We evaluated the trained classifier on a set of held-out spectrograms computed via DBA-based averaging over our test split. As shown in Table 1, DBA produces vowels with spectral characteristics that correspond to the correct harmonic feature sufficiently well for classification.

class	precision	recall	F_1	support
front	0.850	0.752	0.798	218
back	0.793	0.877	0.833	236
accuracy			0.817	454

Table 1: Precision, recall, F_1 score, and accuracy of CNN classifier on held out DBA-generated vowel tokens

5 Conclusions and Future Work

We have presented an exemplar-based production algorithm leveraging *dynamic time warping* and *DTW barycenter averaging* operating over phonetically rich, temporally-variable instance representations, overcoming a core challenge for exemplar production models. The work here represents an initial step toward a fully articulated model of exemplar-based phonetic and phonological competence and performance.

In ongoing work we are investigated means of mitigating the noisiness of MNEMORPHON's outputs via principled reductions in exemplar cloud size e.g. exploiting indexical information such as speaker identity, pitch, rate, as well as distance-based restrictions as in Pierrehumbert's 2001 original approach.

In addition, we believe that the work here opens a path modeling productive morphophonological alternations such as Turkish vowel harmony or English past tense.

data generating process, network architecture, and training procedure.

References

Isaac Elias. 2006. Settling the intractability of multiple alignment. *Journal of computational biology : a journal of computational molecular cell biology*, 13(7):1323–1339.

Anmol Gulati et al. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Keith Johnson. 2007. Decisions and Mechanisms in Exemplar-based Phonology. In Maria-Josep Solé, Patrice Speeter Beddor, and Manjari Ohala, editors, *Experimental Approaches to Phonology*. Oxford University Press.

Robert Kirchner, R. Moore, and T-Y Chen. 2010. Computing phonological generalization over real speech exemplars. *Journal of Phonetics*, 38.

Frédéric Mailhot. 2010. Instance-based acquisition of vowel harmony. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics.

François Petitjean, A. Ketterlin, and P. Gançarski. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3).

Janet B. Pierrehumbert. 2001. Exemplar dynamics: Word frequency, lenition and contrast. *Typological Studies in Language*, 45:1–11.

H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Taras K. Vintsyuk. 1968. Speech discrimination by dynamic programming. *Cybernetics*, 4:52–57.

Özgül Salor, Bryan Pellom, Tolga Çiloglu, Kadri Hacioglu, and Mübeccel Demirekler. 2002. On developing new text and audio corpora and speech recognition tools for the turkish language. In *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 349–352.