Acoustic barycenters as exemplar production targets

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

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

Abstract

We present a solution to the problem of exemplar-based language production from variable-duration tokens, leveraging algorithms from the domain of time-series clustering and classification. Our model stores and outputs tokens of phonetically rich and temporally variable representations of recorded speech. We show qualitatively and quantitatively that model outputs retain essential acoustic/phonetic characteristics despite the noise introduced by averaging, and also demonstrate the effects of similarity and indexical information as constraints on exemplar cloud selection.

1 Introduction

We present here an exemplar production model that implements solutions to the challenges of measuring between-exemplar distance (i.e. alignment) and fostering phonetic generalization over speech tokens of variable duration (Pierrehumbert, 2002; Kirchner et al., 2010).[1] Our model, Mnemorphon, makes use of algorithms for alignment and averaging the domain of time-series clustering and classification. We show qualitatively by direct inspection of model outputs and quantitatively via statistical classification that Mnemorphon’s outputs retain essential acoustic/phonetic characteristics, despite noise introduced by averaging, and also demonstrate the effects constraining exemplar cloud composition by means of similarity weighting and indexical information.

We begin with an overview of exemplar-based approaches to phonetics and phonology, highlighting the core production challenges of temporal variability and generalization. We then introduce Mnemorphon and present our experiments and results, and finish with some discussion of planned work and future directions.[2]

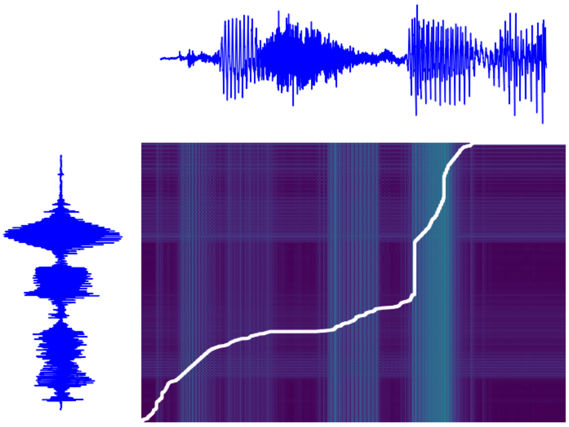

Figure 1: Dynamic time warping alignment of waveforms of two tokens of the Turkish word *kuşları* (*“birds”*), highlighting temporal variability.

2 Exemplar-based phonetics and phonology

Exemplar-based theories of categorization propose that humans classify percepts by direct comparison to memorized exemplars of previous experiences (Semon, 1923; Medin and Schaffer, 1978; Hintzman, 1986; Nosofsky, 1986), whereas linguistic theories have traditionally been couched in terms of symbolic categories that abstract away from details of usage and experience. When experiments in speech perception suggested that human word recognition is facilitated by fine details of remembered experiential episodes, e.g. speakers’ voices (Goldinger, 1996, 1998), phoneticians began to explore the possibilities of memory-based approaches. Johnson (1997a,b) presented a pair of exemplar models of phonetic perception that

[1] Here and below, “length”, “duration”, “variability”, etc. specifically refer to *temporal extent*, rather than e.g. number of phones/segments. Where we discuss discrete sequences, it is assumed that sequence coordinates represent a fixed and constant temporal duration.

[2] Code for the model, experiments, and evaluations described below will be made available at https://github.com/calicolab/mnemorphon.

Proceedings of the 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 67–76
June 20, 2024 ©2024 Association for Computational Linguistics

provided elegant and novel accounts of speaker normalization and speech segmentation. Soon after Pierrehumbert (2001) published the first implemented model of exemplar-based phonological *production*, in the context of a production-perception feedback-loop model of sound change.

These initial investigations ushered in a flurry of subsequent research in exemplar-theoretic phonetics and phonology in areas as diverse as sound change, categorical emergence and entrenchment, sociophonetic variation, frequency effects in productivity, the status of abstract phonetic categories, and the induction of morphophonological alternations (Bybee, 2001; Pierrehumbert, 2001; Hawkins, 2003; Wedel, 2006; Gahl and Yu, 2006; Johnson, 2006; Ettlinger, 2007; Kirchner et al., 2010; Mailhot, 2010a).

Goldrick and Cole (2023) provide a recent overview of the theoretical and empirical successes, along with some outstanding potential challenges, of exemplar-based approaches to production. The core theoretical challenges faced by exemplar-based models of production are handling input variability, particularly with respect to temporal variation, and the need for a mechanism for robust generalization from prior experiences. Below we discuss the first of these, showing how it can be surmounted with a 50 year old approach to speech recognition, and later we address the latter, introducing a 21$^{\text{st}}$ century algorithm for averaging time series.

2.1 The problem of temporal variability

It is well-known that distinct utterances of human speech[3] categories such as words can vary significantly in duration, both within and across speakers (see e.g. Figure 1). This temporal variability is one of the core challenges for any exemplar model. These models typically compute a distance or similarity function over exemplars; we therefore require a means of computing such a measure that is robust to length-wise variation. Fortunately, such an algorithm already exists and is well-known in the speech recognition and time series analysis literatures.

Dynamic time warping (DTW) (Vintsyuk, 1968; Sakoe and Chiba, 1978; Mueen and Keogh, 2016, for a recent overview) is an algorithm for computing a distance measure between sequences of potentially differing lengths. Given a pair of sequences X, Y with *coordinates*[4] $[x_0, ..., x_n], [y_0, ..., y_m]$ embedded in a shared parametric space D^k and a distance function $d(x_i, y_j)$, DTW finds the best alignment between X and Y via the following optimization:

$$DTW(X,Y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2} \qquad (1)$$

Here π is an alignment or *warping path* between X and Y; a sequence of pairs $((i_1, j_1), ..., (i_k, j_k))$ each of whose elements respectively indexes positions in X and Y, with the following properties: (i) $\pi_1 = (1,1)$ and $\pi_k = (n,m)$, that is, the start and end of X and Y are aligned, (ii) π increases monotonically in i and j, and (iii) each $i \in [1, ..., n]$ and $j \in [1, ..., m]$ appears at least once in π. The *DTW distance* between X and Y is the minimized sum of coordinate-wise distances over all possible alignments.

We note here that we are not the first to realize that DTW provides a solution to the problem of temporal alignment in exemplar production; Kirchner et al.'s (2010) PEBLS incorporates it in modeling a phonological generalization on toy data using speech tokens. Their approaches requires ad-hoc modifications to the DTW algorithm, along with an additional hierarchical clustering step to mitigate the problem of spurious generalizations (see Appendix A for a more detailed overview). Below we examine a more principled approach to the latter problem.

2.2 Generalizing production from exemplar knowledge

As alluded to above, a remaining challenge for production exemplar models is accounting for the human capacity to "go beyond the data" and generalize over prior experiences, a hallmark of cognition. In exemplar models of perception/comprehension, a distance measure and simple nearest-neighbour search are sufficient to enable generalization; given an input form, the listener finds the previously stored form that is closest to it in the representational space, and assigns that form's category to the input form, then stores them together.

[3] We focus on speech here and below, but believe the approach developed here applies, *mutatis mutandis*, to signed languages as well.

[4] We borrow this term from Petitjean et al. (2011) for individual (possibly multidimensional) elements of sequences and use it throughout.

In production, the speaker has a given category and must produce an output for it. The simplest means of doing so is to select a previously-memorized token from within that category and directly produce it. This method effectively turns the model into a look-up table, making it in-principle incapable of generalizing beyond the input to which it has been exposed (consider whether this approach could handle e.g. a "wug" test Berko Gleason, 1958). We turn now to one means of surmounting this obstacle.

Pierrehumbert (2001) presents a model of phonological production that implements generalization via a simple but ingenious method of exemplar composition. The model's exemplars are points in $(F1, F2, F3)$ formant space, representing vowel steady-state measurements, paired with vowel category labels. For a given vowel category $\mathcal{C}$, generation of an output exemplar c_{out} proceeds in three steps: (i) a single *seed* token c_{in} is randomly selected from all stored exemplars associated with $\mathcal{C}$, (ii) an analogical set or *exemplar cloud* — C_{in} is constructed by considering all exemplars within a fixed Euclidean distance of c_{in} in formant space, and finally (iii) an output token c_{out} is produced by computing a *similarity-weighted* average of the exemplar cloud, with similarity computed as an inverse exponential function of distance.

Many phonetic and phonological insights have been derived from exemplar models that take inspiration from this approach, averaging over point-like data in low-dimensional spaces (e.g. Wedel, 2006; Ettlinger and Johnson, 2010, *inter alia*). This approach can be straightforwardly extended to handle parametric spaces of higher dimensionality e.g. encoding richer acoustic information with spectral frames, or sociophonetic context such as interlocutor identity, etc. However, it is unclear how it might be extended to incorporate the *dynamic* nature of human language, which unfolds in time and cannot be reduced to point measures. That is, the problem of straightforwardly accommodating the temporal variability and generalization of human speech in implemented production models remains underexplored.[5]

3 Mnemorphon: A bit of progress in exemplar-based production[6]

Any implemented exemplar model must minimally include tokens of some primitive linguistic unit encoded in a suitable representational format, associated category labels, and a means of computing analogically relevant similarity between exemplars (Johnson, 2007). For Pierrehumbert's model discussed above, these are segments (specifically vowels), the space defined by tuples of the first three formants, and inverse Euclidean distance in formant space. For Mnemorphon these architectural parameters are as follows:

- **Units:** tokens are complete words, with no representation of sub-lexical linguistic categories (syllables, segments, etc.)
- **Representation:** exemplars are encoded as mel-scaled spectrograms (Deng and O'Shaughnessy, 2003)
- **Categories:** each exemplar is associated to a discrete "lexical" label encoded as a pseudo-phonemic character string for mnemonic convenience, roughly corresponding to a word meaning
- **Similarity:** similarity between tokens is computed as an inverse function of DTW distance (see below for details)

Our general task can now be framed as follows: given a seed exemplar and a cloud of tokens of possibly varying lengths from a given category, we seek a procedure by which we can generate an output exemplar as an "average" of the cloud.

As it happens, exactly computing the sample mean of a set of sequences with potentially differing lengths corresponds to solving the problem of *multiple sequence alignment*, which is known to be computationally intractable (Elias, 2006). Notwithstanding this, there are tractable approximation methods that are theoretically justifiable and empirically suitable; Petitjean et al. (2011) introduce one such approach, *DTW barycenter averaging* (DBA).

3.1 Computing averages of variable-length sequences

DBA is an algorithm that takes as input a set of sequences and iteratively converges to an average sequence that is locally optimal, in the sense of

[5] To our knowledge Kirchner et al. (2010)is the only extant model to address it to date.

[6] With apologies to Goodman (2001).

Algorithm 1 DBA (adapted from Petitjean et al., 2011)

```
Require: $\mathcal{S}$ the sequences to average
Require: $\hat{s} = [\hat{s}_1 ... \hat{s}_k]$ initial barycenter
  converged ← False
  assocTbl ← table of length $k$
  while converged $\neq$ True do
    for all $s \in \mathcal{S}$ do
      $\pi \leftarrow$ DTW($\hat{s}, s$)
      for all $(i, j) \in \pi$ do
        assocTab[$i$] ← assocTab[$i$] $\cup$ $s_j$
    for all $\hat{s}_i \in \hat{s}$ do
      $\hat{s}_i \leftarrow$ BARYCENTER(assocTab($i$))
    CHECKCONVERGENCE
  return $\hat{s}$
```

minimizing a quantity analogous to *inertia* in k-means clustering (i.e. "within-cluster variance" MacQueen, 1967):

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

The sequence $\hat{s}^*$ is called the *barycenter* of the set of sequences $\mathcal{S}$, by analogy with the use of that term for *center of mass*, a dynamical physical points which need not equal or intersect with any of the points it averages over.

Given a set of sequences $\mathcal{S}$ and an initial "best-guess" barycenter $\hat{s}$ (typically randomly generated or sampled directly from $\mathcal{S}$), DBA iterates over two phases (see Algorithm 1):

- **Align:** compute DTW alignments for $\hat{s}$ and each $s \in \mathcal{S}$, and for each coordinate $\hat{s}_i$ of the barycenter, store the set of all coordinates it was aligned with for each s
- **Update:** update each coordinate $\hat{s}_i$ of $\hat{s}$ to be the barycenter of its associated coordinates found in the alignment phase

The algorithm halts after a predetermined number of iterations, or when the difference in inertia across iterations falls below a preset convergence threshold. 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. In either case, the inertia stays the same or decreases, hence DBA is guaranteed to converge.

Algorithm 2 MNEMORPHON output generation

```
Require: $\Lambda$ a lexicon of categories and associated exemplars
Require: $\lambda \in \Lambda$ the category for which MNEMORPHON must generate an output
  $S_{in} \leftarrow$ GETALLEXEMPLARS($\lambda$)
  $s_i \leftarrow RandomSelectOne(C_{in})$          ▷ the seed
  $C_{in} \leftarrow$ CONSTRUCTCLOUD($S_{in}$)
  $\hat{s}^* \leftarrow$ DBA($C_{in}$, $s_i$)) return $\hat{s}^*$
```

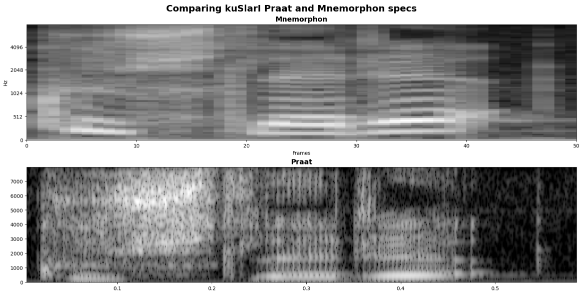

Figure 2: Spectrograms of a token of *kuşları* ("birds"), as created with our parameters versus Praat's default values.

3.2 Generating outputs

With its representations and averaging procedure in place, MNEMORPHON's basic algorithm for exemplar output generation is straightforward (see Algorithm 2):

1. Given a set of stored *(exemplar, category)* pairs, and a target output production category
2. select a *seed* exemplar associated with the target category
3. construct an analogical set or *cloud* from the remaining exemplars in the target category
4. output the mean of the cloud, computed via DBA[7]

We leave the cloud construction step in 3 unspecified here; Pierrehumbert uses a fixed-radius neighbourhood of the seed, but alternatives are possible, e.g. a fixed number of seed neighbours. Below we partially address this question; ultimately it is a model parameter to be tuned empirically.

4 Data

For the experiments described below our raw data set is an audio corpus of Turkish speech, consisting of microphone recordings (16KHz sample rate)

[7] MNEMORPHON uses the implementation of DBA available in the `tslearn` Python package (Tavenard et al., 2020).

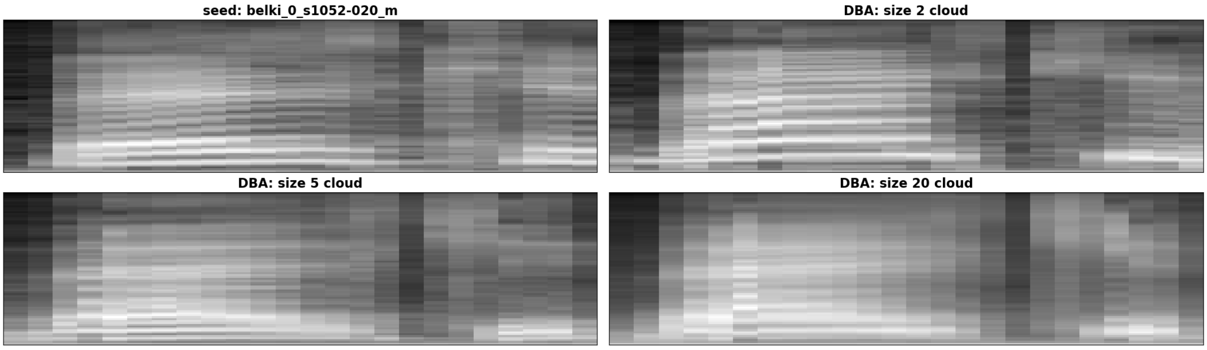

Figure 3: Output spectrogram of seed token of *belki* (“maybe”), along with spectrograms generated from 2, 5, and 20 tokens. Cloud size correlates with noisy outputs.

from 120 speakers (balanced across binarized gender categories; age 19–50 years, mean=23.9) who each read 40 sentences sampled from a triphone-balanced set of 2462 Turkish sentences (Özgül Salor et al., 2006). Metadata for each speaker includes (binarized) gender, dates of birth and recording, places of birth and residence, and level of education. 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.

Each recorded sentence is transcribed in standard Turkish orthography as well as an ASCII-compatible phonemic orthography derived from SAMPA (Wells, 1997), called *METUbet* (Özgül Salor et al., 2002). The corpus also includes word-level, phone-level, and HMM state alignments, computed with an HMM-GMM acoustic model trained on a subset of the full set of sentences.

As with most linguistic corpora, word frequencies follow a roughly Zipfian distribution. There are 7412 words in our dataset, the most frequent of which, *bir* (“one/a”), occurs approximately 897 times, whereas there are 2423 words which occur only once.

4.1 Model inputs

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 illustrated in Figure 2, these spectrogram parameters generate comparatively coarse *narrow-band* spectrograms, unlike e.g. Praat’s default values which have finer temporal resolution and are perhaps better suited to visual presentation. Our choice of spectrogram parameters was constrained by our evaluation methodologies, discussed below.

5 Experiment 1: cloud composition

In our initial experiments we explore the effect of cloud composition on Mnemorphon’s outputs. We begin with a maximally unconstrained approach, conditioning cloud selection solely on word category membership. For each of the word categories (i.e. distinct METUbet strings) represented in our corpus, we uniform randomly select one token as the seed exemplar and sample progressively larger uniform random subsets of the remaining tokens from the category as the cloud from which Mnemorphon computes a barycenter. We illustrate the outcome here in Figures 3 and 4 for a representative example, the form *belki* (gloss: “maybe”, corpus freq: 40, rank: 43). We plot full spectrograms and a selection of mel bands, respectively, for the seed token along with from Mnemorphon’s output barycenter for varying cloud sizes.

Figure 3 shows clearly that increasing the number of tokens included in the cloud results in Mnemorphon’s output spectrograms becoming “blurrier”, losing most of the fine structure present in individual tokens, particularly with respect to frequency information. Notwithstanding this noise, the individual mel bands plotted in Figure 4 show that Mnemorphon’s generation algorithm

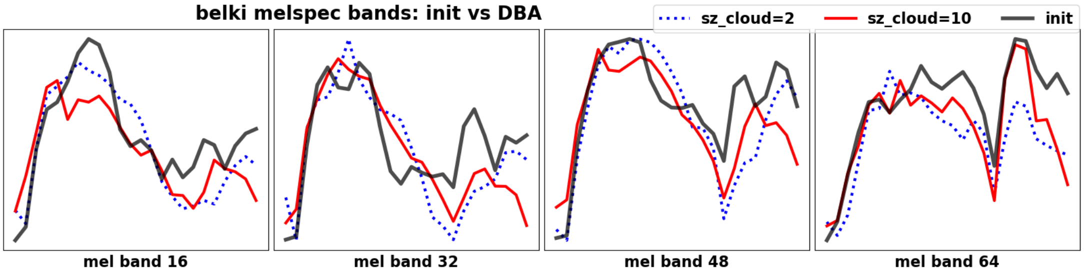

Figure 4: Mel bands 16, 32, 48, 64 of seed spectrogram, with DBA spectrograms from cloud size 2, and 10, for seed token *belki*.

does find meaningful averages for temporally variable signals, locating and aligning the major peaks and troughs in the energy for each band along the temporal dimension.

The relation between output noise and increasing numbers of cloud exemplars found here is not solely due to cloud size, but rather that the clouds' tokens are dispersed in the parametric space. To confirm this we generate outputs from the same seed, this time with two small clouds of the same size (N=3), constrained to contain the maximally similar and dissimilar tokens in the category, respectively. The outputs, shown in Figure 5, confirm that dispersion plays a key role in the quality of MNEMORPHON's generated forms. This in turn raises the question of latent categorical structure or organization within exemplars clouds.[8]

As discussed in Section 2, some of the early motivation for exploration of exemplar-based speech processing was the apparent storage and use of *non-linguistic* information, for example indexical information. The results above suggest that constraining MNEMORPHON's cloud selection by using any such additional contextually available information would likely serve to further reduce the output variance, resulting in cleaner, and in a sense more representative, output spectrograms. To test this we re-ran the same experiment as above, with the same seed token, this time constraining MNEMORPHON's clouds by using the (binarized) gender information that is available in our corpus. We used the output from our initial, unconstrained, experiment for a cloud size $N = 10$, and then used ten uniform randomly selected tokens from the relevant category that were tagged `F` ("female") in our corpus. Once again, we see in Figure 6 that constraining MNEMORPHON's cloud along dimensions of similarity, linguistic or otherwise, yields cleaner, more representative outputs.

Notwithstanding the obscuring or blurring of phonetic detail in MNEMORPHON's outputs, larger scale patterns of energy distribution across different frequency bands and time slices remain visible, hinting at an emergent, transient form of abstraction; a hallmark of exemplar models. In our next experiment we see that there is indeed linguistic categorical information recoverable from these outputs.

In addition to the direct visual evaluation here, we use a publicly available pre-trained neural vocoder (Lee et al., 2023)[9] to re-synthesize audio from our generated spectrograms for impressionistic auditory evaluation.[10] It is the use of this vocoder that constrained the spectrogram parameters in our data preparation; because BigVGAN is trained on narrow-band spectrograms (the standard choice in neural text-to-speech synthesis), these are required for any subsequent synthesis. That said, the finer frequency resolution of narrow-band spectrograms is likely beneficial for the quantitative evaluation in Experiment 6.

6 Experiment 2: latent categorical information in MNEMORPHON's outputs

We have seen that MNEMORPHON's outputs quickly become noisy as a function of cloud size, although this is somewhat mitigated by heavily weighting the influence of cloud tokens that are close to the seed in DTW distance. Despite this noise, we wish to determine whether generated outputs retain any categorically characteristic phonetic signal. We investigate this in the present experiment, in which we train a neural network to take spectral slices as inputs and classify them as *front* or *back* vowels.

[8] We thank an anonymous reviewer for highlighting this point and encouraging us to explore it.

[9] `https://github.com/NVIDIA/BigVGAN`

[10] The accompanying website hosts samples of audio synthesized from MNEMORPHON's outputs.

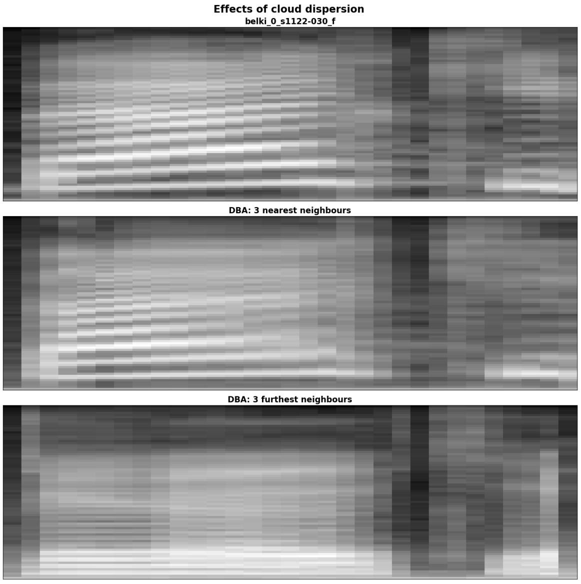

Figure 5: Effect of cloud dispersal; spectrogram of seed token of *belki*, along with spectrograms generated from clouds with minimal and maximal dispersion (nearest and furthest 3 neighbours, respectively).

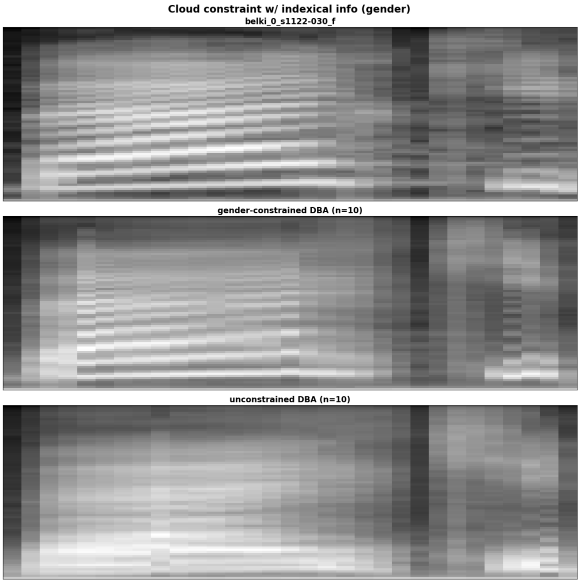

Figure 6: Effect of "gender"-based cloud constraint; spectrogram of seed token of *belki*, along with spectrograms generated from size $N = 10$ clouds restricted to tokens tagged as "female" versus sampled uniformly across gender markers.

Our focus on this particular phonetic characteristic, foreshadows work in progress assessing Mnemorphon's ability to learn productive morphophonological generalizations, in particular Turkish front/back vowel harmony.[11]

Although Mnemorphon itself has no notion 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 converted them directly to mel spectrograms, 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 current 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 layer), a max-pooling layer, and a final fully-connected layer projecting to a binary output (modeling [± back]). Kernel sizes, learning rate and batch size were tuned on a development split; the final training run was for 25 epochs.[12]

6.1 Data augmentation

Like all supervised learning approaches, neural networks are sensitive to distribution shift, where the properties that the network learns to extract as relevant features are differently distributed in the training and evaluation sets. This exact situation obtains in the current experiment, where 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.

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.

6.2 Results

Table 1 shows the precision, recall, and F_1 score of our classifier on the test split of our data set. We can see that Mnemorphon is, at least according to our classifier, producing output vowels with phonetic characteristics that enable their identification as front or back.

[11] Turkish also has rounding harmony, which we also leave for future investigation.

[12] See the accompanying repository for fuller details of the data generating process, network architecture, and training procedure.

class	precision	recall	F_1	support
front	0.880	0.878	0.879	3154
back	0.866	0.867	0.867	2856
accuracy			0.873	6010

Table 1: Precision, recall, F_1 score, and accuracy of CNN phone classifier on held-out set

7 Discussion

We have shown here that *dynamic time warping* and *DBA barycenter averaging* together constitute a viable basis for a production algorithm in an exemplar model, MNEMORPHON, whose token representations are word-sized mel spectrograms of variable durations, overcoming a core challenge for exemplar production models. We showed both qualitatively and quantitatively that despite noise introduced by averaging over tokens that are dispersed in spectrotemporal space, our model's outputs retain phonetic properties that are characteristic of the exemplars from the generating categories.

8 Limitations and future work

MNEMORPHON's production algorithm as applied in these experiments generates comparatively noisy outputs, unless the selection of tokens for the exemplar cloud is severely constrained. Nonetheless, we see this work as an initial step toward a fully articulated theory and model of exemplar-based (psycho)linguistic knowledge. An eventual goal is to assess how far such a "pure" or "core" model can take us before a hybrid approach becomes necessary (cf. Goldrick and Cole, 2023).

In future work will explore further restrictions on cloud construction, exploring e.g. speaker identity, dialect, and speech rate among others.

As hinted in Section 6, we also intend to extend this work to account for productive morphophonological alternations like Turkish vowel harmony (see Mailhot, 2010b, for an exemplar production approach that learns productive vowel harmony on toy data, including patterns of opaque and transparent neutrality), and eventually to data from psycholinguistic research on speech perception and production (e.g. contexts of phonetic reduction and lengthening, and patterns of interlocutor convergence).

As the data used here are not widely accessible, we also intend to reproduce these results in the not-too-distant future using data from the Mozilla Common Voice corpus (Ardila et al., 2020)[13] in order to facilitate reproducibility.

8.1 A note on gender

As a final remark, we acknowledge here that gender identity and expression exist on a spectrum, and hence that the use of binarized gender in the experiment on constraining cloud size is problematic. The experiment was added in response to a pertinent reviewer remark, and in the interest of expediency we used the binarized gender markers that are available in our corpus's metadata. In future work we hope to address this more carefully, either using a wider array of self-reported gender identities, or potentially relying purely on phonetic features, e.g. high or low F_0 (although of course this is at best an approximation).

Acknowledgements

We would like to thank our reviewers for valuable feedback. Although this work has not been directly shared with any audiences, many of the ideas here are the result of hallway discussions and email exchanges with friends and colleagues over several years.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Jean Berko Gleason. 1958. The child's learning of english morphology. *Word*, 14.

Joan Bybee. 2001. *Phonology and Language Use*. Cambridge Studies in Linguistics. Cambridge University Press.

Li Deng and Douglas O'Shaughnessy. 2003. *Speech Processing: A dynamic and optimization-oriented approach*. Marcel Dekker, New York, NY.

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

Marc Ettlinger. 2007. An exemplar-based model of chain shifts. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVI)*.

[13] https://commonvoice.mozilla.org

Marc Ettlinger and Keith Johnson. 2010. Vowel discrimination by english, french and turkish speakers: Evidence for an exemplar-based approach to speech perception. *Phonetica*, 66(4):222–242.

Susanne Gahl and Alan C. L. Yu. 2006. Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review*, 23(3):213–216.

Stephen Goldinger. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 22(5):1166–1183.

Stephen Goldinger. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105:251–79.

Matthew Goldrick and Jennifer Cole. 2023. Advancement of phonetics in the 21st century: Exemplar models of speech production. *Journal of Phonetics*, 99.

Joshua T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Sarah Hawkins. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31(3):373–405. Temporal Integration in the Perception of Speech.

Douglas Hintzman. 1986. "schema abstraction"' in a multiple-trace memory model. *Psychological Review*, 93:411–428.

Keith Johnson. 1997a. The auditory/perceptual basis for speech segmentation. In *OSU Working Papers in Linguistics*, 50, pages 101–113. Ohio State University. Department of Linguistics.

Keith Johnson. 1997b. Speech perception without speaker normalization: an exemplar model. In K. Johnson and J.W. Mullenix, editors, *Talker Variability in Speech Processing*. Academic Press, San Diego.

Keith Johnson. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34(4):485–499. Modelling Sociophonetic Variation.

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.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. BigVGAN: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297.

Frédéric Mailhot. 2010a. 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.

Frédéric Mailhot. 2010b. *Modelling the acquisition and evolution of vowel harmony*. Ph.D. thesis, Carleton University.

Douglas L. Medin and Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Abdullah Mueen and Eamonn Keogh. 2016. Extracting optimal performance from dynamic time warping. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2129–2130, New York, NY, USA. Association for Computing Machinery.

Robert M. Nosofsky. 1986. Attention, similarity, and the context theory of classification. *Journal of Experimental Psychology*, 115:39–57.

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.

Janet B. Pierrehumbert. 2002. Word-specific phonetics. In *Laboratory Phonology 7*, pages 101–140, Berlin, New York. De Gruyter Mouton.

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.

Richard Semon. 1923. *Mnemic Psychology*. George Allen & Unwin, London. Translated by B. Duffy (original work publish 1909).

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. 2020. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6.

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

Andrew Wedel. 2006. Exemplar models, evolution and language change. *The Linguistic Review*, 23(3):247–274.

John Wells. 1997. Sampa - computer readable phonetic alphabet. Accessed on January 20, 2024.

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.

Özgül Salor, Bryan Pellom, Tolga Çiloglu, Kadri Hacioglu, and Mübeccel Demirekler. 2006. Middle east technical university turkish microphone speech (v1.0) ldc2006s33. Web download. Philadelphia; Linguistic Data Consortium.

A PEBLS : Phonological Exemplar-based Learning System

Kirchner et al. (2010) present PEBLS, to our knowledge the only exemplar production model in the phonetics/phonology literature that operates over (digitized representations of) real speech tokens.

To produce an output, PEBLS randomly selects a seed token from the set of word labels; all remaining exemplars in that set serve as the cloud. Output production is then cast as the problem of determining an optimal alignment between the seed and the entire cloud. Concretely, PEBLS's output is a token composed of coordinates or sub-sequences of in-cloud exemplars that may occur in *any position* in *any token*. The optimization is computed over both coordinate-wise similarities, and inter-coordinate transition similarities (these obtained by computing an alignment of the cloud with itself, offset by one coordinate.)

Kirchner et al. note that this production method also faces the issue of generalization, as for categories whose exemplars mostly-with-exceptions reflect some phonological generalization (e.g. intervocalic lenition). If the initially sampled seed token violates the relevant generalization (i.e., it includes a stop between vowels), and even a single generalization-violating exemplar exists in the cloud, it will be directly output by PEBLS, notwithstanding the preponderance of generalization-conforming exemplars.

In order to predispose PEBLS to produce tokens that reflect the statistical generalizations instantiated in its exemplars, a "confidence" measure is introduced that expresses the representativeness of sequences of coordinate transitions within the cloud. This confidence computation requires a complete hierarchical clustering over the cumulative partial DTW scores at each coordinate transition.

While PEBLS presents solutions to the problems of production and generalization over real speech exemplars, it does so at the cost of non-trivial complexity; introducing unmotivated modifications to the DTW algorithm, along with an adhoc mechanism to down-weight the importance of non-representative exemplars within a cloud.