Modeling speech recognition and synthesis simultaneously: Encoding and decoding lexical and sublexical semantic information into speech with no access to speech data

Anonymous ACL submission

Abstract

Human speakers encode information into raw speech which is then decoded by the listeners. This complex relationship between encoding (production) and decoding (perception) is often modeled separately. Here, we test how decoding of lexical and sublexical semantic information can emerge automatically from raw speech in unsupervised generative deep convolutional networks that combine both the production and perception principle. We introduce, to our knowledge, the most challenging objective in unsupervised lexical learning: an unsupervised network that must learn to assign unique representations for lexical items with no direct access to training data. We train several models (ciwGAN and fiwGAN by Beguš 2021) and test how the networks classify raw acoustic lexical items in the unobserved test data. Strong evidence in favor of lexical learning emerges. The architecture that combines the production and perception principles is thus able to learn to decode unique information from raw acoustic data in an unsupervised manner without ever accessing real training data. We propose a technique to explore lexical and sublexical learned representations in the classifier network. The results bear implications for both unsupervised speech synthesis and recognition as well as for unsupervised semantic modeling as language models increasingly bypass text and operate from raw acoustics.

1 Introduction

Speech technology has traditionally been divided into two parts: automated speech recognition (ASR) and speech synthesis. Hearing humans, however, perform both tasks — speech production and speech perception with a high degree of mutual influence (the so-called production-perception loop; Wedel 2004; Vihman 2015). Human speakers encode information into raw acoustic speech stream and decode it using both articulation and audition.

This paper proposes that the two principles should be modeled simultaneously and argues that a GAN-based model called ciwGAN/fiwGAN (Beguš, 2021) learns linguistically meaningful representations from both production and perception. In fact, lexical learning in the architecture emerges precisely from the requirement that the network for production and the network for perception interact and generate data that is mutually informative. We show that with only the requirement to produce informative data, the models not only produce desired outputs (as argued in Beguš 2021), but also learn to classify lexical items in a fully unsupervised way from raw unlabeled speech.

1.1 Prior work

Computational models of lexical learning from speech data have a long history (Räsänen, 2012). Earlier work operated with pre-assumed features that needed to be extracted from acoustic speech stream. Recent models operate directly from acoustic speech stream and involve a variety of modeling approaches, from Bayesian to neural modeling (Levin et al., 2013; Lee et al., 2015; Chung et al., 2016; Chrupała et al., 2017; Shafaei-Bajestan and Baayen, 2018; Kamper, 2019; Chorowski et al., 2019; Baayen et al., 2019). A recent push towards unsupervised learning from raw audio means that models of lexical learning are cognitively more plausible (Levin et al., 2013; Kamper et al., 2014; Chung et al., 2016; Hu et al., 2020; Baevski et al., 2020; Niekerk et al., 2020; Beguš, 2021) as hearing infants learn words primarily from unlabeled acoustic speech stream.

Most of the existing models of lexical learning, however, focus primarily on either ASR/speech-to-text (perception) or text-to-speech/speech synthesis (production; see Wali et al. 2022 for an overview). Variational Autoencoders (VAEs) involve both an encoder and decoder, which allows unsupervised acoustic word embedding as well as generation of speech, but these proposals only use VAEs for either unsupervised ASR (Chung et al., 2016; Chorowski et al., 2019; Baevski et al., 2020; Niekerk et al., 2020) or for speech synthesis/transformation (e.g. Hsu et al. 2017). Earlier neural models replicate brain mechanisms behind perception and production (Tourville and Guenther, 2011; Guenther and Vladusich, 2012), but they do not focus on lexical learning or classification and do not include recent progress in performance of deep learning architectures. GAN-based synthesizers are mostly supervised and get text or acoustic features in their input (Kumar et al., 2019; Kong et al., 2020; Bińkowski et al., 2020; Cong et al., 2021). Donahue et al. (2019) propose a WaveGAN architecture, which can generate any audio in an unsupervised manner, but does not involve a lexical classifier — only the Generator and the Discriminator, which means the model only captures synthesis and not classification (the same is true for Parallel WaveGAN; Yamamoto et al. 2020). Beguš (2021) proposes the first textless fully unsupervised GAN-based model for lexical representation learning, but evaluates only the synthesis (production) aspect of their model by only evaluating outputs of the Generator network.

1.2 New challenges

Here, we model lexical learning with a classifier network (the Q-network) that mimics perception and lexical learning and is, crucially, trained from another network's production data (the Generator network). Using this architecture, we can *both* generate new words in a controlled causal manner by manipulating the Generator's latent space as well as classify novel words from unobserved test data withheld from training in a fully unsupervised manner.

This paper also introduces some crucial new challenges to the unsupervised acoustic word embedding and word recognition paradigm (Dunbar et al., 2017, 2019, 2020). First, the architecture requires extremely low vector representation of lexical items. In the fiwGAN architecture, the network needs to represent 2^n of classes with only n variables. To our knowledge, no other proposal features such dense representation of acoustic lexical items.

Second, the models introduce a challenge to learn meaningful representations of words without ever accessing training data. The lexical classifier network is twice removed from training data. The Q-network learns to classify words only from the Generator's outputs and never accesses training data. But the the Generator never accesses the training data either – it learns to produce words only by maximizing the Discriminator's error rate. This means that the classifier needs to learn to represent unique lexical items in a highly challenging setting, where training data is two levels removed — only the Discriminator actually accesses training data.

Finally, lexical learning in the proposed architecture is fully unsupervised. VAEs are a prominent architecture in the unsupervised lexical learning paradigm. The encoder-decoder architecture learns representations of lexical items in an unsupervised manner, but the generation principle is not unsupervised: the decoder (equivalent to the Generator) is trained on generating data such that the distance between input data and the generated output is minimized. In other words, the decoder has full access to the input data and needs to replicate it. In the GAN framework, on the other hand, the Generator needs to learn to produce lexical items for each unique code; the classifier needs to learn to assign unique code to each lexical item by only accessing the Generator's outputs. Neither the Generator nor the Q-network have access to training data and they do not replicate input data, which means they are fully unsupervised both in the generation and in the classification aspect.

Why are these challenges important? First, representation learning with highly reduced vectors is more interpretable and allows us to analyze the causal effect between individual latent variables and linguistically meaningful units in the output of the synthesis/production part of the model (Section 4.3). We also can examine the causal effect between linguistically meaningful units in the classifier's input and the classifier's output in the perception/recognition part of the model (Section 4.2.2).

Reduced vectors also enable analysis of the interaction between individual latent variables. For example, each element (bit) in a binary code (e.g., [1, 0], [0, 1], [1, 1]) can be analyzed as a feature ϕ_n (e.g. [ϕ_1, ϕ_2]). Such encoding allows both holistic representation learning and featural representation learning. We can test whether each unique code corresponds to unique lexical semantics and how individual features in binary codes ([ϕ_1, ϕ_2]) interact/represent sublexical information (e.g. presence of a phoneme; Section 4). Reduced vectors also allow us to model lexical semantics similarly with established methods of computational semantic modelling, where meaning is often represented with binary codes (such as in Steinert-Threlkeld and Szymanik 2020).

Second, humans acquire speech production and perception with a high degree of mutual influence (Vihman, 2015). The production-perception loop can facilitate speech acquisition (e.g. in L2 learning; Baese-Berk 2019). Production and perception also make language dynamic and cause change over time (Ohala, 1993). Modeling production (synthesis) and perception (recognition) simultaneously will help us build more dynamic and adaptive systems of human speech communication that are closer to reality than current models which treat the two components separately. This can be beneficial both to speech technologies and to cognitive models of speech acquisition.

Third, the paper tests learning of linguistically meaningful representations in one of the most challenging training settings. Results from such experiments test the limits of deep learning architectures for speech processing. This paper argues that learning of linguistically meaningful representations self-emerges even in these highly challenging learning conditions.

Fourth, unsupervised ASR (Baevski et al., 2021) and "textless NLP" (Lakhotia et al., 2021) have the potential to enable speech technology in a number of languages with no or little resources. Many of these languages feature substantially richer phonological systems than English. Most deep generative models for unsupervised learning focus exclusively on either lexical (see above) or phonetic learning (Eloff et al., 2019; Shain and Elsner, 2019) and do not model phonological learning. The fiwGAN architecture with its featural latent space allows simultaneous holistic lexical representation learning and sublexical learning of phonological contrasts. Exploring how the two levels interact will be increasingly important as speech technology becomes available to phonologically rich languages.

Finally, speech technology is shifting towards unsupervised learning (Baevski et al., 2021). Our understanding of how biases in data are encoded in unsupervised models is even more poorly understood than in supervised models. The paper proposes a way to test how linguistically meaningful units self-emerge in fully unsupervised models for word learning, how contrasts are encoded in the latent space, and how they interact with other variables in a classifier network for unsupervised ASR. Speech carries a lot of potentially harmful social information (Holliday, 2021); a better understanding of how linguistically meaningful units self-emerge and get encoded and how they interact with other features in the data is the first step towards mitigating the risks of unsupervised deep generative ASR models.

2 Models

We use Categorical InfoWaveGAN (ciwGAN) and Featural InfoWaveGAN (fiwGAN) architectures (Beguš 2021; based on WaveGAN in Donahue et al. 2019 and InfoGAN in Chen et al. 2016). In short, the ciwGAN/fiwGAN models each contain three networks: a Generator G that upsamples from random noise z and a latent code c to audio data using 1D transpose convolutions, a Discriminator D that estimates the Wasserstein distance between the generated output $G(z, c)$ using traditional 1D convolutions, and a Q-network Q that aims to recover c given Generator output $G(z, c)$. As in the traditional GAN framework, the Generator and Discriminator operate on the same loss in a zero-sum game, forcing the Generator to create outputs similar to the training data. However, the Generator (along with the Q-network) is additionally trained against the accuracy of the Q-network, forcing the Generator to maximize the mutual information between latent code c and generated output $G(z, c)$ and the Q-network to recover this relationship. CiwGAN models c as a one-hot vector of several classes, while fiwGAN models c using a binary encoding.

In other words, the Generator learns to produce outputs that resemble speech from latent space (that includes both code c variables and random z variables) without direct access the training data. The Q-network takes generated audio speech data ($G(z, c)$) and needs to figure out what code c the Generator used for each particular output. The Generator needs to encode unique code vector c such that the Q-network will be successful in retrieving unique information (unique code vector c) only from the Generator's audio outputs. Lexical learning thus needs to emerge in a fully unsupervised manner, only from the requirement of the Generator to produce informative data. The training data contains unlabeled raw acoustic data. The Generator could in principle encode any information into

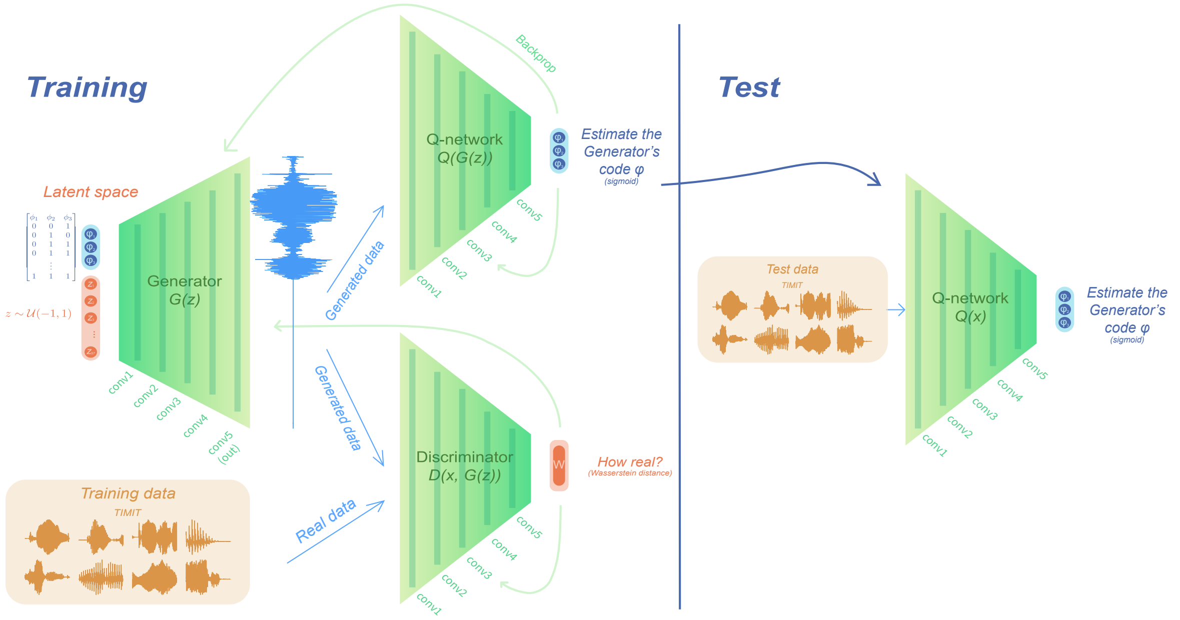

Figure 1: Architecture of the fiwGAN network (Beguš, 2021) as used during the training and test tasks.

the code variables, but given the structure of the data, the most informative way to encode information is to associate each lexical item with a unique code.

Previous work on ciwGAN and fiwGAN (Beguš, 2021) has focused on the ability of the Generator to learn meaningful representations of c that encodes phonological processes and lexical learning, with no exploration of the Q-network. In this paper, we focus on the Q-network's propensity for lexical learning. Towards this end, we maintain the architecture of a separate Q-network (in contrast to the original InfoGAN proposal, where Q is estimated by appending additional hidden layers after the convolutional layers of the Discriminator). This allows us to explore a fully unsupervised lexical classifier network that takes lexical items in raw audio form and classifies them with unique codes (Figure 1).

3 Experiments

To simultaneously test the performance of the production (synthesis) and perception (classification) in the Generator and the Q-network, we train three networks: one using the one-hot (ciwGAN) architecture on 8 lexical items from TIMIT, one with binary code (fiwGAN) architecture on 8 lexical items from TIMIT. To test how the proposed architecture scales up to larger corpora, we also train a fiwGAN network on 508 lexical items from LibriSpeech (Panayotov et al., 2015).[1]

3.1 Data

The lexical items used in 8-words models are: *ask*, *dark*, *greasy*, *oily*, *rag*, *year*, *wash*, and *water*. A total of 4,052 tokens are used in training (approximately 500 per each word). The words were sliced from TIMIT and padded with silence into 1.024s .wav files with 16kHz sampling rate which the Discriminator takes as its input.

In the LibriSpeech experiment, 508 words were chosen. We discarded the 78 most common lexical items in the Librispeech train-clean-360 dataset (Panayotov et al., 2015), because of their disproportionate high frequency (5,290 to 224,173 tokens per word). We then arbitrarily choose the 508 next most common words for training, resulting in a total of 757,120 tokens. The individual counts for each word in the training set ranges from 571 to 5,113 tokens.

3.2 Perception/classification

To test if the Q-network is successful in learning to classify lexical items without ever accessing training data, we take the trained Q-network from the architecture (in Figure 1) and feed it novel, unobserved data. In other words, we test if the Q-network can correctly classify novel lexical items by assigning each lexical item a unique code.

[1] Trained checkpoints and data will be released.

Altogether 1,067 test data in raw waveforms from unobserved TIMIT were fed to the Q-network (both in the ciwGAN and fiwGAN architectures). The raw output of this experiment are pairs of words with their TIMIT transcription and the unique code that the Q-network outputs in its final layer. We test the performance of the models using inferential statistics rather than comparison to existing models due to the lack of models with similarly challenging learning objectives.

To perform hypothesis testing on whether lexical learning emerges in the Q-network, we fit the word/code pairs to a multinomial logistic regression model using the *nnet* package (Venables and Ripley, 2002). The dependent variable are the input words (8 classes), the independent variable is the final code c that the Q-network outputs for each input. To test lexical learning, we compare Akaike Information Criterion (AIC; Akaike 1974) of a model with code c as predictor and an empty model. In the ciwGAN setting (one-hot encoding), AIC of a model with c as a predictor is substantially lower ($2129.056, df = 56$) than the empty model ($4448.191, df = 7$). Figure 2 gives predicted values for each code/lexical item. The figure suggests that most lexical items have a clear and substantial rise in estimates for a single unique code. This suggest that the Q-network learns to classify novel unobserved TIMIT words into classes that correspond to lexical items.

Lexical learning emerges in the binary encoding (fiwGAN) as well, but the code vector is even more reduced in this architecture (3 variables total), which makes error rates higher compared to the ciwGAN architecture (Figure 2). The multinomial regression model with unique binary codes as predictors fits the data better than the empty model ($\text{AIC} = 3248.071, df = 56$ for the model with the predictor and $\text{AIC} = 4448.191, df = 7$ for the empty model).

Estimates of the fiwGAN multinomial regression models in Figure 2 suggest most unique codes feature a single clear and substantial raise in regression estimates for each unique lexical item. Words like *rag* or *wash* do not seem to have a clear learned representation in the latent code c.

The binary nature of the latent code in the fiwGAN architecture allows testing of whether individual features (ϕ_1, ϕ_2, ϕ_3) carry lexical information. All individual features are significant as separate predictors (according to AIC).

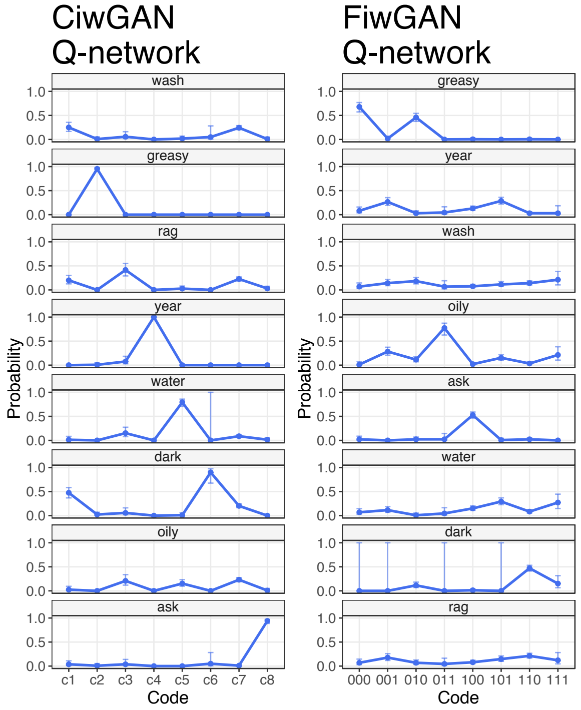

Figure 2: Estimates of a multinomial regression model.

3.3 Production/synthesis

To test the production (synthesis) aspect of the model, we generate 100 outputs for each unique latent code c both in the ciwGAN and fiwGAN setting (1,600 outputs total). According to Beguš (2020); Beguš (2021), setting latent codes to marginal values outside of the latent space reveals the underlying value of each latent code, which is why we generate data with code variables set at 3 (e.g. [0, 0, 3], [0, 3, 3], etc). One hundred outputs per each code for each model (ciwGAN and fiwGAN) were analyzed by a compensated trained phonetician who was not a co-author on this paper. The annotator annotated generated outputs as either featuring the eight lexical items, deviating from the eight items (annotated as *else*), or as unintelligible outputs (also *else*).[2]

Figure 3 illustrates lexical learning in the Generator network. Code variables are significant predictors of generated words according to the AIC test in both models. The learned representations are very similar both in the Q-network and in the Generator. One advantage of the Generator network is that we can force categorical or near categorical out-

[2] The following regex coding of annotations was used: if "^wa(sh|tch)$", then *wash*, if "[^s]e[ae]r", then *year*, if "water.*$", then *water*, if "[ao][wia]l[iy].*$" then *oily*, if "rag.*$", then *rag*, if "dar.*$", then *dark*, if "greas.*$", then *greasy*, if "as.*$", then *ask*.

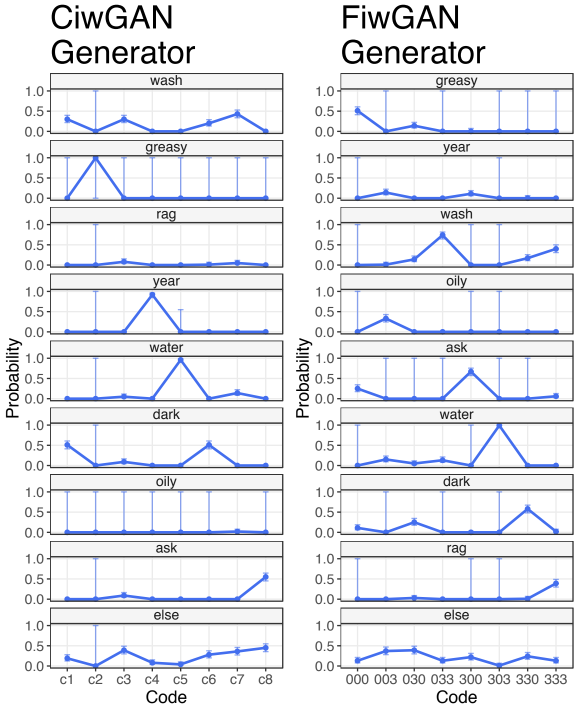

Figure 3: Estimates of a multinomial regression model.

puts by manipulating latent variables to marginal values outside of training range (e.g. in our case to 3). For example, *greasy* has 100% success rate in ciwGAN; *water* 99% in fiwGAN and 96% in ciwGAN.

4 Holistic and featural learning

Binary encoding allows simultaneous holistic lexical encoding (unique code = lexical item) as well as featural learning, where features (bits) correspond to sublexical units such as phonemes (e.g. [s] or [ʃ]). This paper proposes a technique to explore lexical and sublexical learned representations in a classifier network. To test whether evidence for sublexical learning emerges in the perception aspect of the proposed model, we annotate inputs to the Q-network for any sublexical property and use regression analysis with each feature as a predictor to test how individual features correspond to that property.

4.1 TIMIT

We focus on one of the the most phonetically salient sublexical properties in the training data: presence of a fricative [s], [ʃ]. We include the word for *dark* among the words containing [s] because a high proportion of *dark* tokens feature [s] frication (due to *dark* standing before *suit* in TIMIT). The data were fit to a logistic regression linear model with presence of [s] in the input test data as the dependent variable and the three features (ϕ_1, ϕ_2, ϕ_3) as predictors. Estimates of the regression model in Figure 4 suggest that the network encodes a sublexical phonemic property (presence of frication noise of [s]) with $\phi_3 = 0$.

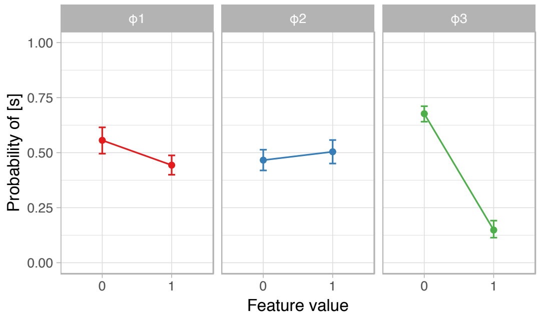

Figure 4: Estimates of a logistic regression model.

4.2 LibriSpeech

To test how the proposed technique of unsupervised lexical and sublexical learning extends to larger corpora, we test the Q-network trained on 508 lexical items from LibriSpeech. The model has 9 latent feature variables ϕ which yields $2^9 = 512$ classes. Altogether 10,914 test tokens (withheld from training) of the 508 unique words were fed to the Q-network in fiwGAN architecture trained for 61,707 steps.

4.2.1 Holistic representation learning

First, raw classification of outputs suggest that holistic lexical learning in the Q-network emerges even when the training data contains a substantially larger set (508 items and a total of 757,120 tokens) and a more diverse corpus. The training data here too is twice removed from the Q-network and the test data was never part of the training. Figure 5 illustrates four chosen lexical items and the codes with which they are represented. Each lexical item features a peak in one unique code. To verify that this particular code indeed represents that particular lexical item, we also analyze which other lexical items are classified with the most frequent code for each of the four chosen lexical item. There too, each code represents one lexical item more strongly.

To test how frequent such well-learned representations are, we randomly selected 20 out of the 508 lexical items, which includes items with extremely low representation in the training and test corpora

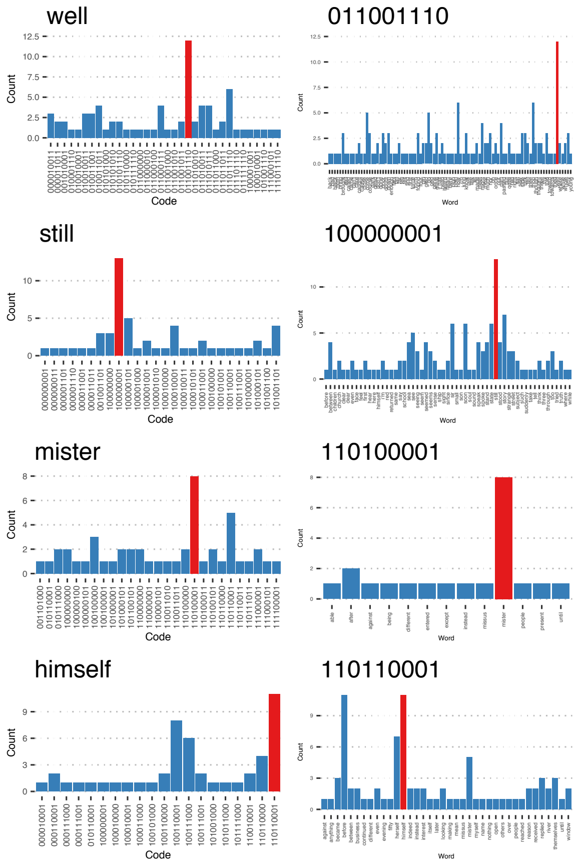

Figure 5: (left) Raw counts of code distribution per each of the four chosen tested word (from unobserved test data). The word with highest count is color-coded in red. (right) Raw counts of word distributions per code that has the highest count for each word (in the left graph).

(e.g. N $=$ 7). Of the 20 randomly selected words, 4 (20%) have representations such that the peak per lexical item corresponds to the peak per the most frequent code (as the first three examples in Figure 5). In 5 further cases (25%), two or more peaks have the same, but not higher counts than the word/code peak pair (for a total of 45% of successful outcomes if both groups are counted as successful). In the remaining 55% (11 items), the peaks do not match across the word/code pairs.[3]

These counts are fully deterministic and therefore conservative. The distribution of code variables per each word are, however, not independent. For example, the second most common code for *mister* in Figure 5 differs from the most common one in only one feature (digit). Violation in a single feature value is equally treated as violation in multiple feature values in our counts.

Likewise, there is substantial amount of phonetic similarity in words classified by a single code. For example, the word most commonly classified with [100000001] is indeed *still*, but other frequent words for this classification code are *state*, *stand*, *stood*, *story*, etc. (Figure 5).

4.2.2 Featural representation learning

These similarities suggest that the network encodes sublexical properties using individual features in the binary code. To quantitatively test this hypothesis, we test how the network encodes presence of word-initial [s] ([#s]). Frication noise of [s] is a phonetically salient property and restricting it to word-initial position allows us to test featural and positional encoding.

Librispeech word/Q-network code pairs are annotated for presence of word-initial [s] (dependent variable) and fit to a logistic regression linear model with the nine feature variables ϕ_{1-9} (part of the binary code) as independent predictors. Regression estimates are given in Figure 6. Three features (ϕ_2, ϕ_3, and ϕ_5) correspond to presence of initial [s] substantially more strongly than other features. It is reasonable to assume that the network encodes this sublexical contrast with the value of the three features (ϕ_2, ϕ_3, ϕ_5) 0. It would be efficient if the network encodes word-initial [s] with 3 features, because there are approximately 54 s-initial words. The 6 feature codes remaining besides ϕ_2, ϕ_3, ϕ_5 allows for $2^6 = 64$ classes.

To verify this hypothesis, the presence of [s] in input words (dependent variable) is fit to a logistic regression model with only one predictor: the value of the three features ϕ_2, ϕ_3, ϕ_5 with two levels: 0 and 1. Only 5.0% [4.6%, 5.5%] of words classified with ϕ_2, ϕ_3, and ϕ_5 = 1 contain word-initial [s], while 47.9% [44.3%, 51.5%] of words classified as ϕ_2, ϕ_3, and ϕ_5 = 0 contain word-initial [s] (see estimates in Figure 6).

4.3 Featural learning in production

The fiwGAN architecture allows us to test both holistic and featural learning in both production and perception. Value 0 for ϕ_2, ϕ_3, ϕ_5 has been associated with word-initial [s] in the Q-network (perception). To test whether the Generator matches the Q-network in this sublexical encoding, we generate sets of outputs in which all other ϕ variables (except ϕ_2, ϕ_3, and ϕ_5) and all z-variables are held constant, but the ϕ_2, ϕ_3, and ϕ_5 variables are interpolated from 0 to 3 in increments of 0.2. We

[3] We counted one case with all counts equal across the word/code pair as unsuccessful.

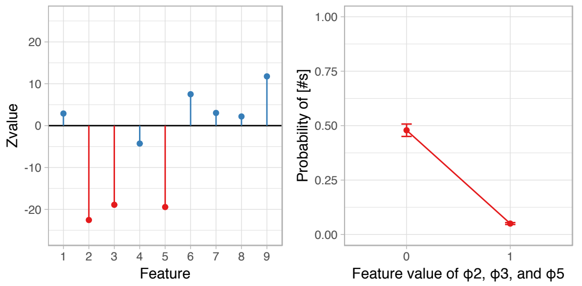

Figure 6: (left) Estimates of a logistic regression model based on the LibriSpeech Q-network with presence of word-initial [s] as the dependent variables and the 9 features in the binary code are the predictors. (right) Estimated probability that the classified word contains a word-initial [s] when the outcome in classification for ϕ_2, ϕ_3, and ϕ_5 are 0 and 1.

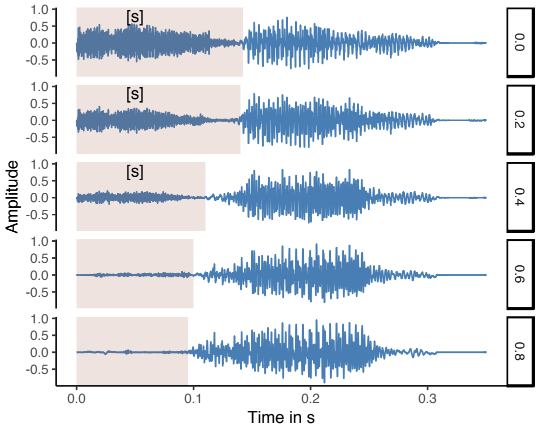

Figure 7: Outputs of the Generator network (waveforms) when ϕ_2, ϕ_3, and ϕ_5 are simultaneously interpolated from 0.0 to 0.8 while all other latent variables are held constant.

analyze 20 such outputs (where the other ϕ variables and z-variables are sampled randomly for each of the 20 sets).

In 11 out of the 20 generated sets (or 55%), word-initial [s] appears in the output for code ϕ_2, ϕ_3, ϕ_5 = 0 and then disappears from the output as the value is interpolated.[4] Additionally, in the majority of these cases (approximately 8), the change from [s] to some other word-initial consonant is the only major change that happens as the output transitions from [s] to no [s] with interpolation. In other words, as we interpolate values of the three features representing [#s], we observe a causal effect in the generated outputs as [#s] gradually changes into a different consonant with other major acoustic properties remaining the same in the majority of cases. Figure 7 illustrates this causal effect: the amplitude of the frication noise of [s] gradually attenuates with interpolation, while other acoustic properties remain largely unchanged. The sublexical encoding of word-initial [s] is thus causally represented with the same code both in the Generator network and in the Q-network.

5 Conclusion

This paper demonstrates that a deep neural architecture that simultaneously models the production/synthesis and perception/classification component learns linguistically meaningful units — lexical items and sublexical properties (such as presence of a sound) — from raw acoustic data in a fully unsupervised manner. Lexical and sublexical learning emerge simultaneously only from the requirement of the Generator to output informative data. In this architecture, we can both (i) generate lexical items in a controlled and predictable manner by manipulating individual variables of the latent space in the Generator network and (ii) classify unobserved test lexical items with a unique highly reduced vector representation in the Q-network.

We introduce several challenges to the unsupervised acoustic word embedding paradigm. These challenges, while increasing difficulty of the learning objective, bring several new insights into the unsupervised speech processing/textless NLP paradigm (Dunbar et al., 2017; Lakhotia et al., 2021). The results of the three computational experiments (8-word TIMIT ciwGAN/fiwGAN and 508-word LibriSpeech fiwGAN) suggest that learning emerges despite these challenges. Highly reduced vector representations enable interpretable semantic exploration of the latent space and exploration of the causal effect between the latent space and generated outputs. We also demonstrate that deep convolutional networks are able to classify raw audio into unique word classes whereby training data is twice removed from the classifier network — the Q-network only learns from the Generator's output (in the production-perception loop) and never accesses training data. Finally, the ability to simultaneously model holistic lexical representation learning (in the form of unique binary codes) and sublexical (phonetic and phonological) representations in the form of individual feature codes will be increasingly important as unsupervised speech processing technology becomes more widely available in languages with rich phonological processes.

[4] Annotated by the authors because presence of [s] is a highly salient feature.

References

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Rolf Harald Baayen, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019.

Melissa Michaud Baese-Berk. 2019. Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, 81(4):981–1005.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*, pages 1–12.

Gašper Beguš. 2021. CiwGAN and fiwGAN: Encoding information in acoustic data to model lexical learning with Generative Adversarial Networks. *Neural Networks*, 139:305–325.

Gašper Beguš. 2020. Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. *Frontiers in Artificial Intelligence*, 3:44.

Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C. Cobo, and Karen Simonyan. 2020. High fidelity speech synthesis with adversarial networks. In *International Conference on Learning Representations*.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc.

Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aäron van den Oord. 2019. Unsupervised speech representation learning using WaveNet autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2041–2053.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622, Vancouver, Canada. Association for Computational Linguistics.

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *Interspeech 2016*, pages 765–769.

Jian Cong, Shan Yang, Lei Xie, and Dan Su. 2021. Glow-WaveGAN: Learning speech representations from gan-based variational auto-encoder for high fidelity flow-based speech synthesis. In *Proc. Interspeech 2021*, pages 2182–2186.

Chris Donahue, Julian J. McAuley, and Miller S. Puckette. 2019. Adversarial audio synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2019. The zero resource speech challenge 2019: TTS without T. In *Proc. Interspeech 2019*, pages 1088–1092.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330.

Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2020. The zero resource speech challenge 2020: Discovering discrete subword and word units. *Interspeech 2020*.

Ryan Eloff, André Nortje, Benjamin van Niekerk, Avashna Govender, Leanne Nortje, Arnu Pretorius, Elan Biljon, Ewald van der Westhuizen, Lisa Staden, and Herman Kamper. 2019. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. In *Proc. Interspeech 2019*, pages 1103–1107.

Frank H. Guenther and Tony Vladusich. 2012. A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, 25(5):408–422. Is a neural theory of language possible? Issues from an interdisciplinary perspective.

Nicole R. Holliday. 2021. Perception in black and white: Effects of intonational variables and filtering conditions on sociolinguistic judgments with implications for ASR. *Frontiers in Artificial Intelligence*, 4:102.

Wei-Ning Hsu, Yu Zhang, and James Glass. 2017. Learning latent representations for speech generation and transformation. In *Proc. Interspeech 2017*, pages 1273–1277.

Yushi Hu, Shane Settle, and Karen Livescu. 2020. Multilingual jointly trained acoustic and written word embeddings. In *Proc. Interspeech 2020*, pages 1052–1056.

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

Herman Kamper, Aren Jansen, Simon King, and Sharon Goldwater. 2014. Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 100–105.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, and Emmanuel Dupoux. 2021. Generative spoken language modeling from raw audio. *CoRR*, abs/2102.01192.

Chia-ying Lee, Timothy J. O'Donnell, and James Glass. 2015. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403.

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

Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. 2020. Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge. *Interspeech 2020*.

John J. Ohala. 1993. Sound change as nature's speech perception experiment. *Speech Communication*, 13(1):155–161.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Okko Räsänen. 2012. Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication*, 54(9):975 – 997.

Elnaz Shafaei-Bajestan and R. Harald Baayen. 2018. Wide learning for auditory comprehension. In *Proc. Interspeech 2018*, pages 966–970.

Cory Shain and Micha Elsner. 2019. Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 69–85, Minneapolis, Minnesota. Association for Computational Linguistics.

Shane Steinert-Threlkeld and Jakub Szymanik. 2020. Ease of learning explains semantic universals. *Cognition*, 195:104076.

Jason A. Tourville and Frank H. Guenther. 2011. The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, 26(7):952–981. PMID: 23667281.

William N. Venables and Brian D. Ripley. 2002. *Modern Applied Statistics with S*, fourth edition. Springer, New York. ISBN 0-387-95457-0.

Marilyn Vihman. 2015. Perception and production in phonological development. In *The Handbook of Language Emergence*, chapter 20, pages 437–457. John Wiley & Sons, Ltd.

Aamir Wali, Zareen Alamgir, Saira Karim, Ather Fawaz, Mubariz Barkat Ali, Muhammad Adan, and Malik Mujtaba. 2022. Generative adversarial networks for speech processing: A review. *Computer Speech & Language*, 72:101308.

Andrew Wedel. 2004. Category competition drives contrast maintenance within an exemplar-based production/perception loop. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, pages 1–10, Barcelona, Spain. Association for Computational Linguistics.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel Wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203.