Graphone Model Interpolation and Arabic Pronunciation Generation

T. Li, P. C. Woodland, F. Diehl & M. J. F. Gales

Department of Engineering, University of Cambridge,
Trumpington Street, Cambridge, CB2 1PZ, U.K.
{tl284,pcw,fd257,mjfg}@eng.cam.ac.uk

Abstract

This paper extends *n*-gram graphone model pronunciation generation to use a mixture of such models. This technique is useful when pronunciation data is for a specific variant (or set of variants) of a language, such as for a dialect, and only a small amount of pronunciation dictionary training data for that specific variant is available. The performance of the interpolated *n*-gram graphone model is evaluated on Arabic phonetic pronunciation generation for words that can't be handled by the Buckwalter Morphological Analyser. The pronunciations produced are also used to train an Arabic broadcast audio speech recognition system. In both cases the interpolated graphone model leads to improved performance.

Index Terms: pronunciation generation, graphone models, Arabic speech recognition

10.21437/Interspeech.2011-613

1. Introduction

N-gram joint-multigram models applied to grapheme-to-phoneme (G2P) conversion tasks are referred to as *n*-gram *graphone* models [1], where the term *graphone* denotes joint grapheme-phoneme units. *N*-gram graphone models are a data-driven approach for learning a set of variable-length units from a training pronunciation dictionary, and an associated *n*-gram sequence model, which can then be used for pronunciation generation.

In recent years, *n*-gram graphone models have been used in pronunciation generation tasks for several languages, including English, German and French [1, 2, 3]. In these graphone-based pronunciation generation tasks, an *n*-gram graphone model was usually trained on a large general dictionary of entries. However, in certain situations there might be variants of the language that have rather different statistical properties to the bulk of the training data. This might for instance occur in dealing with dialect or foreign words for which both the graphemic and phonemic sequences may be rather different to those seen in training. In this case, it might be appropriate to train a separate *n*-gram graphone model for these atypical words; although if this results in being able to use only a small amount of training data, it may perform poorly.

An alternative approach to address the lack of training data is to combine one or more specific graphone models trained on a small amount of e.g. dialect data with a well-trained model estimated on a large dictionary for standard words and pronunciations. This paper investigates an approach in which a mixture of *n*-gram graphone models is used for this problem, such that the probability assigned to a particular graphone sequence is a linear interpolation of the probabilities from the component models. This is directly analogous to the use of mixtures of *n*-gram models in statistical language model adaptation [4].

The mixture *n*-gram graphone technique is applied to Arabic pronunciation generation in this paper. In Arabic, the short vowels (*fatha* /a/, *kasra* /i/ and *damma* /u/) and diacritics (*shadda*, *sukun*) are commonly not marked in texts. Additionally, *nunation* can result in a word-final *nun* (/n/) being added to nouns and adjectives in order to indicate that they are unmarked for definiteness. Furthermore there are typically many different legal pronunciations for each graphemic sequence which generally correspond to different meanings.

The standard approach to generate pronunciations for Arabic phonetic systems is to use the Buckwalter Morphological Analyser (BMA) [5]. While BMA can analyse and generate pronunciations for the majority of words (here denoted "BW" words) encountered in a large vocabulary broadcast audio transcription system, typically about 25% of the words cannot be processed since the graphemic forms don't follow the rules of Modern Standard Arabic (MSA). This typically includes generating the pronunciations for names, foreign words and dialect words, referred to here as "NBW" words. Since we have access to a small amount of hand-labelled pronunciation data which includes NBW words, a separate *n*-gram graphone can be trained for this data and that model combined with a model trained on a pronunciation dictionary generated by BMA.

In the following sections of this paper, the mathematical basis of a mixture *n*-gram graphone model is first described. Two grapheme-to-phoneme conversion performance measure schemes are then introduced for evaluating pronunciation generation performance with multiple pronunciations. The experimental data, setup, results and discussion for Arabic pronunciation generation are discussed and the impact in terms of word error rate on an Arabic speech recognition task described.

2. Graphone model interpolation

2.1. *N*-gram graphone models

Under the *n*-gram graphone framework, after a dictionary entry's grapheme string O and phoneme string Φ are jointly segmented into pairs of grapheme-phoneme joint sequences, i.e. graphones, by a possible *co-segmentation* ψ, the two observable parallel streams (O, Φ) are regarded as the concatenation of the graphones $g_j = (s_j, \gamma_j)$ in the resulting joint string $G = (O, \Phi, \psi) = \{g_1, g_2, \cdots, g_j, \cdots, g_{L_\psi}\}$. Graphone models restrict the variable length of a grapheme sequence s_j to μ ranging from μ_{min} to μ_{max} and the variable length of a phoneme sequence γ_j to ν ranging from ν_{min} to ν_{max}, denoted (μ_{min},μ_{max})(ν_{min},ν_{max}). $\Psi = \{\psi\}$ can be used to rep-

This work was in part supported by DARPA under the GALE program via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

Copyright © 2011 ISCA

resent the set of all possible co-segmentations of (O, Φ) into graphones. Assuming each graphone depends on its preceding $n-1$ ($n > 1$) graphones in the resulting joint string caused by a particular co-segmentation ψ, the probability of ψ for (O, Φ) into L_ψ graphones can be calculated by

$$P(O, \Phi, \psi) = \prod_{j=1}^{L_\psi} P(g_j|h_j)$$

where $h_j = \{g_{j-n+1}, \cdots, g_{j-2}, g_{j-1}\}$ denotes the corresponding *n*-gram graphone history of a maximum length of *n* − 1 graphones if available. The total likelihood of (O, Φ) is computed as the sum over all the possible co-segmentations

$$\mathcal{L}(O, \Phi) = \sum_{\Psi} P(O, \Phi, \psi)$$

For an *n*-gram graphone model trained on a dictionary $\mathcal{D}$ of M unique words which are respectively paired with their corresponding multiple phonetic pronunciation variants $\mathcal{D} = \{\mathcal{W}_1, \mathcal{W}_2, \cdots, \mathcal{W}_m, \cdots, \mathcal{W}_M\}$, $\mathcal{W}_m = \{(O_1^m, \Phi_1^m), (O_2^m, \Phi_2^m), \cdots, (O_i^m, \Phi_i^m), \cdots, (O_{V_m}^m, \Phi_{V_m}^m)\}$, the likelihood of $\mathcal{D}$ can be computed by the following equation when every dictionary entry in $\mathcal{D}$ is treated independently

$$\begin{aligned}
\mathcal{L}(\mathcal{D}) &= \prod_{m=1}^{M} \mathcal{L}(\mathcal{W}_m) \\
&= \prod_{m=1}^{M} \prod_{i=1}^{V_m} \mathcal{L}(O_i^m, \Phi_i^m) \\
&= \prod_{m=1}^{M} \prod_{i=1}^{V_m} \sum_{\Psi_i^m} P(O_i^m, \Phi_i^m, \psi_i^m) \\
&= \prod_{m=1}^{M} \prod_{i=1}^{V_m} \sum_{\Psi_i^m} \prod_{j=1}^{L_{\psi_i^m}} P(g_j|h_j)
\end{aligned}$$

It is noted that the co-segmentation ψ for (O, Φ) into graphones is an unobserved variable, thus maximum likelihood (ML) estimation of the parameter set $\{P(g_j|h_j)\}$ with probabilities of all possible graphone *n*-grams, which define an *n*-gram graphone model, can be performed through the expectation maximization (EM) algorithm [1].

2.2. *N*-gram graphone mixture model

A mixture model results from the linear interpolation of K component *n*-gram graphone models. The graphone *n*-gram probability is then given by

$$P(g_j|h_j) = \sum_{k=1}^{K} \lambda_k P_k(g_j|h_j)$$

where λ_k is the interpolation coefficient of the kth component model. The estimation approach for the interpolation coefficients is based on maximizing the likelihood of the held-out target dictionary data, $\mathcal{L}(\mathcal{D})$; here $\mathcal{D}$ is the held-out data for interpolation tuning. Thus the optimal interpolation coefficient for the kth combined *n*-gram graphone model is calculated by

$$\hat{\lambda}_k^{\mathrm{ML}} = \arg\max_{\lambda_k} \left\{ \prod_{m=1}^{M} \prod_{i=1}^{V_m} \sum_{\Psi_i^m} \prod_{j=1}^{L_{\psi_i^m}} \sum_{k=1}^{K} \lambda_k P_k(g_j|h_j) \right\}$$

under the constraints $0 \leqslant \lambda_k \leqslant 1$ and $\sum_{k=1}^{K} \lambda_k = 1$. In an similar way to find the coefficients in an *n*-gram mixture model for language modelling, the EM algorithm can be used to perform the ML estimation to find the optimal values [6].

For a two-component mixture model (as used in the experimental evaluation in this paper), we simply have

$$P(g_j|h_j) = \lambda P_A(g_j|h_j) + (1-\lambda) P_B(g_j|h_j) \tag{1}$$

3. G2P conversion performance metrics

An Arabic word usually has multiple phonetic pronunciation variants. This makes it difficult to choose proper metrics to measure Arabic grapheme-to-phoneme conversion performance while comparing and evaluating multiple generated pronunciation hypotheses (*n* hypotheses, $n > 1$) against corresponding multiple reference pronunciations (*m* reference pronunciations, $m > 1$) for each test word. This *n-to-m* case is much more complex to evaluate than the *n-to-1* and *1-to-m* cases. For the problem mentioned above, we have used two schemes to measure *n-to-m* grapheme-to-phoneme conversion performance in terms of the *phoneme error rate* (PER) and the *word error rate* (WER).

3.1. Scheme 1: Average

Given an Arabic grapheme-to-phoneme conversion test set with K words, the phoneme error rate for the Average scheme is computed by

$$\%PER = 100 \times \frac{1}{K} \sum_{k=1}^{K} \frac{\frac{1}{M_k} \sum_{i=1}^{M_k} (I_i^k + D_i^k + S_i^k)}{N_k^{\mathrm{ave}}}$$

where I_i^k, D_i^k and S_i^k respectively denote the number of phoneme insertions, deletions and substitutions when the phoneme sequence of the ith generated pronunciation in total M_k hypotheses of the kth test word is aligned to the reference pronunciation variant which has the smallest Levenshtein Distance to this ith hypothesis among all the reference pronunciations of the kth word, and N_k^{ave} represents the average of the phoneme numbers of all the reference pronunciations for the kth word. The corresponding word error rate is given by

$$\%\,WER = 100 \times \frac{1}{K} \sum_{k=1}^{K} \frac{1}{M_k} \sum_{i=1}^{M_k} \delta_i^k$$

where the indicator variable

$$\delta_i^k = \begin{cases} 0 & \text{if the } i\text{th generated pronunciation} \\ & \text{exactly matches any one of the reference} \\ & \text{pronunciations for the } k\text{th test word,} \\ 1 & \text{otherwise.} \end{cases}$$

3.2. Scheme 2: Best-Match

Given the same test set as above, the phoneme error rate and the word error rate for the Best-Match scheme are respectively calculated by

$$\%PER = 100 \times \frac{1}{K} \sum_{k=1}^{K} \frac{\min_{i \in [1, M_k]} (I_i^k + D_i^k + S_i^k)}{N_k^{\mathrm{ave}}}$$

$$\%\,WER = 100 \times \frac{1}{K} \sum_{k=1}^{K} \min_{i \in [1, M_k]} \delta_i^k$$

where M_k, N_k^{ave}, I_i^k, D_i^k, S_i^k and δ_i^k have the same meaning as above.

4. Experiments on Arabic data

The experimental evaluation of the techniques described above considered the situation where two *n*-gram graphone models are to be combined in a mixture model. One graphone model is estimated on a background training dictionary generated by BMA (version 2) for words that it can analyse (BW words), and the other separate model is generated on a small hand-labelled set of names, foreign words and dialect words (NBW words). These two graphone models are interpolated and the single parameter λ estimated as the mixture component weight for the NBW-trained model.

4.1. Pronunciation generation setup

For this work training (& test) pronunciations for the NBW model were obtained from Arabic Treebank[1] which contains hand-labelled pronunciations. The data in total contains 5,465 NBW words with a total of 7,912 pronunciations. This NBW data was split into a training set (4,947 words with 7,151 pronunciations) and a test set (518 words with 761 pronunciations). To train a graphone model for the BW data, the remaining 80K words from the Treebank sources[2] were passed through BMA to yield 143K pronunciations. This data was split into a training set of 70K words and a 10K test set.

We developed *n*-gram graphone models for automatic pronunciation generation using our locally-developed training tools. Initial experiments showed that shorter span graphones combined with a long-span *n*-gram graphone model yielded the best performance for Arabic pronunciation generation in both the BW and the NBW cases. Hence graphone models employed (1,1)(1,2) graphone length constraints. Training of the individual *n*-gram graphone models used the EM algorithm, and a maximum n-gram order of four was used.

A uniform distribution was used to initialise the graphone training, with higher-order models incrementally built initialised with the lower level model. Thus *n*-gram graphone models were respectively initialized using their own previous $(n-1)$-gram model. A *fractional absolute discounting* [7] scheme was used to address the sparseness problem in the graphone models.

The interpolation coefficient λ for the NBW model in Equation 1, the only tuning parameter, was fixed from the unigram to the four-gram interpolated models. It was found that an optimal value (measured on the test data) was 0.9, and it was confirmed that this was consistent if only half the test set was used to estimate this value (i.e. a held-out set was used to estimate λ). As can be seen in Figure 1, a large difference in log likelihood between this interpolated model and the NBW-only model ($\lambda = 1.0$) shows the benefit of the interpolated approach.[3]

In generating pronunciations for NBW words from the interpolated model, the most likely graphone sequences and the corresponding phoneme sequences are found. The posterior probability of each graphone sequence was used as the pronunciation probability of the corresponding inferred pronunciation for the given word. The number of pronunciation variants for each NBW word in the Treebank source was examined and it

[1]The LDC catalog numbers of the 4 text data are LDC2005T02, LDC2005T20, LDC2005T30 and LDC2004T0, and the 4 acoustic data GALE-Y1Q1, GALE-Y1Q2, GALE-Y1Q3 and GALE-Y1Q4 .

[2]The distributed Treebank pronunciations for these BW words were consistent with those generated by BMA.

[3]Note that the log likelihood of the BW-only model gave a much poorer log likelihood of -38.6.

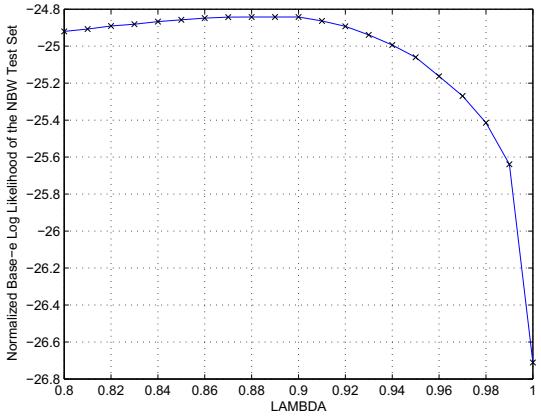


Figure 1: *Normalized base-*e *log likelihood of the NBW test set for λ ranging from 0.8 to 1 with step 0.01.*

was found that there were 96.7% words the pronunciation variant number of which was no more than five. Therefore the five most likely pronunciation variants and their normalised pronunciation probabilities were retained as generated dictionary entries for each test word.

4.2. Pronunciation generation results

The performance of each of the component graphone models as well as the interpolated model was evaluated on both the BW and NBW test sets, and the results given in Table 1. The performance of the BW-trained graphone model ($\lambda = 0$) on the BW test set showed very good performance. However the performance of this system on the NBW test set was poor: the WER using the Best-Match scoring was over 80%. For the NBW-trained graphone system ($\lambda = 1$) the performance on the BW test set was poor, whereas the performance on the NBW test set, though not as good as the BW-trained graphone model on the BW test set, was far better than the BW-trained graphone model. This clearly indicates that the pronunciation statistics for the BW and NBW words are very different.

Test set	λ	Average		Best-Match	
		PER	WER	PER	WER
BW	0	6.63	40.78	0.82	5.09
	0.9	36.48	91.45	24.72	68.64
	1	73.05	99.96	63.32	99.52
	Union	14.44	74.81	2.66	17.20
NBW	0	46.25	96.41	29.58	83.98
	0.9	20.68	82.08	3.38	18.34
	1	20.15	83.32	4.17	21.24
	Union	28.42	84.86	7.83	31.47

Table 1: *PER (%) and WER (%) for the BW and NBW test sets when λ is equal to 0, 1 and 0.9. Also shown is the performance of a single graphone model "Union" trained directly on the union of the BW and NBW training sets.*

The performance of the interpolated graphone model ($\lambda = 0.9$) was evaluated on the BW and NBW test sets. The results are shown in Table 1. As expected the performance on the BW test set was poorer than the BW model ($\lambda = 0$). However this is

not an issue as in practice BMA will be used to generate these pronunciations. On the NBW test set with the Best-Match scoring, the use of the interpolated graphone model showed reductions in both PER and WER, yielding better performance than either of the individual component models. The performance of a further model trained on the union of the BW and NBW training dictionaries (denoted "Union" in Table 1) shows that the interpolated model is clearly superior on the NBW test set.

4.3. Speech recognition experiments

This section focused on the further application of the best interpolated four-gram graphone model ($\lambda = 0.9$) to NBW word pronunciation generation for an Arabic speech recognition task. The pronunciation generation process was the same as above. For comparison, the Arabic pronunciation generation method introduced in [8], the *rule-based* method for short here, was used as the baseline method for the same tasks. For this rule-based method, a set of single letter pronunciation rules were derived from a dictionary generated by the Buckwalter analysis. After splitting all words into letter sequences, acoustic training data were then force-aligned using these pronunciations. Next the alignment result was evaluated and the pronunciation probability for each pronunciation rule was estimated. Based on these pronunciation probabilities, the top 5 most likely pronunciations for a given word were generated using the rules.

Two phonetic systems, V1 and V2, were built on about 175 hours of acoustic model (AM) training data released by LDC under the DARPA GALE programme. The word list of this AM training data included 82,847 BW words, the multiple pronunciations of which were obtained from BMA, and 12,556 NBW words which need pronunciations from the two methods described above. The rule-based and the interpolated four-gram graphone-based pronunciations of these NBW words were respectively used in producing phone-level training transcripts for the V1 and V2 acoustic models. A corresponding graphemic system G0 was also trained on the same data. The graphemic and the two phonetic systems respectively used 36 graphemes and 39 phones. State-clustered trigrapheme and triphone Hidden Markov Models (HMMs) were used in G0 and V1/V2 maximum likelihood (ML) model training. LMs were trained on 22 Arabic sources including broadcast news, broadcast conversation, newswire, and web data. The vocabulary size for all LMs was 350K words.

The experiments in this section made use of the multi-pass decoding framework for Arabic described in [8]. The graphemic system G0 was used for the P1 and P2 stages in the initial experiments which were used for lattice generation and adaptation supervision generation. The third P3 decoding stage used adapted versions of either the V1 or V2 models. For P3 decoding, the rule-based method and the interpolated four-gram graphone model were respectively used to generate pronunciations for about 97K NBW words in the 350K word list; pronunciations of the rest of BW words were obtained from the Buckwalter analysis.

The word error rate performance was measured on the combination of 9 GALE test sets from between 2006 and 2009 that contain both broadcast news and broadcast conversation data. These sets contains a total of 25 hours of data and about 179K test tokens. In total the *in-vocabulary* (IV) NBW word tokens are a very small proportion in each test set (0.79%-3.25%), and over all 9 test sets averages 1.56%. We therefore extracted the test segments with at least one IV NBW word token from these 9 test sets ("NBW segments") to highlight the effect of IV NBW word pronunciations on recognition results. The proportion of IV NBW word tokens in the NBW segments set was 5.42%. We further examined the error rate on just the IV NBW words in the reference transcription to further focus on the effect of the modified dictionary.

System	Overall	NBW segments	NBW tokens
V1	20.41	27.21	28.80
V2	20.29	27.00	25.40

Table 2: *WER (%) for the V1 and V2 systems.*

Table 2 shows the P3 decoding results contrasting the use of rule-based and interpolated graphone-based NBW word pronunciations in both the AM training and the system testing. The results show that the interpolated graphone-based method results in a small improvement in word error rate of on average about 0.1% absolute. While this is a small change, there is a consistent improvement across all test sets used. However when focusing on just the NBW segments, there is a larger reduction in WER of about 0.2% absolute and on the IV NBW word tokens there is a reduction in error rate of 3.4% absolute. Hence while the overall impact of the interpolated graphone model on WER performance is small, it has a far greater impact on the NBW words.

5. Conclusions

In this paper a mixture *n*-gram graphone model was described. This interpolates probabilities from separate graphone models and was applied to the problem of generating pronunciations for Arabic words that can't be handled by the Buckwalter Morphological Analyser. Experimental results on Arabic tasks suggest that the technique is useful for automatic pronunciation generation and produces improved speech recognition performance on the words with revised pronunciations. While the experiments were performed on a fairly small training data setup, the method was used in training our recently developed state-of-the-art Arabic transcription system [9].

6. References

[1] Bisani, M. and Ney, H., "Joint-sequence models for grapheme-to-phoneme conversion", *Speech Communication*, vol. 50, no. 5, pp. 434-451, 2008.

[2] Wang, D., King, S., Frankel, J. and Bell, P., "Stochastic pronunciation modelling and soft match for out-of-vocabulary spoken term detection", in *Proc. ICASSP*, 2010.

[3] Badr, I., McGraw, I. and Glass, J., "Learning new word pronunciations from spoken examples", in *Proc. Interspeech*, 2010.

[4] Bellegarda, J. R., "Statistical language model adaptation: review and perspectives", *Speech Communication*, vol. 42, no. 1, pp. 93-108, 2004.

[5] Available at http://www.ldc.upenn.edu.

[6] Kneser, R. and Steinbiss, V., "On the dynamic adaptation of stochastic language models", in *Proc. ICASSP*, 1993.

[7] Tam, Y. C. and Schultz, T., "Correlated bigram LSA for unsupervised language model adaptation", in *Proc. Neural Information Processing Systems (NIPS)*, 2008.

[8] Diehl, F., Gales, M. J. F., Tomalin, M. and Woodland, P. C., "Phonetic pronunciations for Arabic speech-to-text systems", in *Proc. ICASSP*, 2008.

[9] Diehl, F., Gales, M. J. F., Tomalin, M. and Woodland, P. C., "Word boundary modelling and full covariance Gaussians for Arabic speech-to-text systems", in *Proc. Interspeech*, 2011.