Integrating Conditional Random Fields and Joint Multi-gram Model with Syllabic features for Grapheme-to-Phone Conversion

Xiaoxuan Wang, Khe Chai Sim

School of Computing, National University of Singapore

wangxx@comp.nus.edu.sg, simkc@comp.nus.edu.sg

Abstract

In this paper, we present a hybrid system that combines the Joint Multi-gram Model (JMM) and the Conditional Random Field (CRF) classifiers to solve the Grapheme-to-Phone (G2P) conversion problem. JMM is a generative language model for the n-grams of the joint letter-phoneme units. JMM is able to model longer phonetic contextual information. However, it is difficult to incorporate complex features, such as syllabification structures, to JMM. On the other hand, CRFs can be used to perform G2P by formulating the task as a sequence-labeling problem. CRFs are discriminative classifiers that can incorporate complex feature functions. However, modeling in CRFs requires the alignment between the letters and phonemes. Furthermore, traditional linear chain CRFs usually only employ bigram output information for practical reasons, which is not sufficient for this task. In this work, JMM and CRFs are combined in tandem to yield the JMM-CRF hybrid system that benefits from both of the individual approaches. Results on the CMUDict and CELEX databases show that the proposed hybrid system consistently outperforms the individual JMM and CRF systems. Finally, syllabic features are incorporated into the CRFs as additional features and achieve further performance improvement with the hybrid system.

Index Terms: grapheme-to-phone conversion, conditional random fields, joint multi-gram model, speech recognition

1. Introduction

Grapheme-to-phone conversion (G2P) refers to predicting pronunciation of a word given its orthography. It has important applications in human language technologies, especially in speech synthesis, speech recognition and open vocabulary spoken term detection. Many different approaches have been explored by researchers for the G2P conversion. Sejonwski [1] applied neutral network for G2P on NETtalk. The input of the network is a context window of plus/minus three letters. Hain [2] used a feedforward neural network for G2P conversion process and propose a simple algorithm to generate the G2P matching database with a phonetic dictionary as input. In decision tree approach [3], the tree is grown by testing a set of binary questions for each node, and choosing the best question according the information criterion. Hidden Markov Model [4] is employed for G2P task, in which graphemes are observations and states are phones to form a Markov chain.

More recently, a theoretically stringent probabilistic framework called Joint-Multigram Model (JMM) is proposed by Bisani and Ney [5]. In this model, graphemes and phones are modelled together via their joint probabilities. For each word, the letter sequence and its corresponding pronunciation form a joint sequence. A language modelling technique is applied on the joint sequences. As such, JMM is a generative model. Another popular approach is to employ discriminative classifiers, such as Conditional Random Fields (CRFs) [6, 7, 8], for G2P. Since CRFs maximize the conditional probabilities between graphemes and phones directly, G2P could be formulated as a sequential labelling problem. More recently, Jouvet [9] compared JMM and CRFs for multiple pronunciation predication in speech recognition. They join the pronunciation prediction lists from JMM-based and CRFs-based converters together as ASR lexicons. The results show that the training process is robust to some errors in pronunciation lexicon, whereas pronunciation lexicon errors are harmful in the decoding process. The joint prediction lexicon works better than using any of these two system independently. This may indicate the complimentary modelling capabilities between JMM and CRFs.

In this paper, we propose combining the JMM and CRF classifiers to yield a hybrid G2P system. Specifically, multiple letter-phoneme alignments extracted from the JMM lattices are used by CRF to perform G2P conversion. The union of the lattices generated by CRF are then composed with the JMM lattice to produce the final lattice, from which the best phoneme sequence can be obtained. The final lattice encodes a weighted combination of the JMM and CRF scores, where the weights can be empirically adjusted to optimize development data performance. Moreover, motivated by the finding in [10] that syllabification information can help improve the G2P performance, we further investigate the effectiveness of incorporating a set of syllabic features into CRFs.

The structure of the paper is as follows. Firstly, the theoretical framework of JMM and CRFs are introduced in Section 2. Section 3 presents the proposed JMM-CRF hybrid system for G2P. Section 3 describes the CRFs features, including the new sets of features derived from syllabic information. Experimental results based on the CMUDict and CELEX databases are reported in Section 5. We evaluate each individual models and the hybrid system. We also test the effectiveness of syllabic features in this section. Finally, Section 6 summarizes the results and gives the conclusions.

2. Grapheme-to-phone Conversion

2.1. Joint Multi-gram Model

The fundamental idea of Joint Multi-gram Model (JMM) is to model the joint probability of both the letter and phoneme sequences by considering all possible letter-phoneme alignments. According to the JMM formulation, the optimum phoneme sequence is obtained using the following Bayes' decision rule:

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q} \in Q} P(\mathbf{g}, \mathbf{q} | \theta_{\texttt{JMM}}) \quad (1)$$

where $\mathbf{g}$ and $\mathbf{q}$ denote the letter and phoneme sequences, respectively. $\theta_{\texttt{JMM}}$ are the JMM parameters. $\hat{\mathbf{q}}$ is the best phoneme sequence obtained form a set of possible phoneme sequences, Q. JMM treats the segmentation (or alignment) of the letter and phone sequences as hidden variable. Therefore, the joint probability, $P(\mathbf{g}, \mathbf{q})$, is determined by summing over all matching joint units sequences:

$$P(\mathbf{g}, \mathbf{q}) = \sum_{\mathbf{u} \in S(\mathbf{g}, \mathbf{q})} P(\mathbf{u}) \tag{2}$$

where $S(\mathbf{g}, \mathbf{q})$ represents all possible segmentations; $\mathbf{u} = u_1, u_2, \ldots, u_K$ denotes a sequence of joint units. Each join unit represents a pair of letter and phoneme strings. Therefore, the probability of $\mathbf{u}$ can be modeled using a standard n-gram language model approximation:

$$P(\mathbf{u}) \simeq \prod_{j=1}^{K} P(u_j | u_{j-1}, \ldots, u_{j-n+1}) \tag{3}$$

In this work, we constrained the letter-phoneme alignments such that each joint unit maps zero or one letter to zero or one phoneme. That is to say, one phone (letter) can only be mapped to at most letter (phone).

2.2. Conditional Random Fields

Conditional Random Fields (CRFs) model the conditional probability distribution of the label sequence given an observation sequence [11]. Due to its probabilistic interpretation and discriminative characteristics, CRFs are widely adopted in many natural language processing [12] and spoken language processing [13] tasks. In this work, we treat the G2P conversion as a sequential labelling problem. That is, given the word spelling, $\mathbf{g}$, the optimum phonetic pronunciation, $\hat{\mathbf{q}}$, is predicted as follows:

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q} \in Q} P(\mathbf{q} | \mathbf{g}, \theta_{\texttt{CRF}}) \tag{4}$$

where $\theta_{\texttt{CRF}}$ denotes the CRF model parameters. Formally, a first-order linear-chain CRFs, which assumes the first-order Markov property, defines the conditional probability of a label sequence $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ given input $\mathbf{g} = (g_1, g_2, \ldots, g_m)$ as in Eqn. 5:

$$P(\mathbf{q} | \mathbf{g}; \lambda) = \frac{1}{Z(\mathbf{g}, \lambda)} \exp\left(\sum_m \lambda_m F_m(\mathbf{g}, \mathbf{q}) \right) \tag{5}$$

$F_m(\mathbf{g}, \mathbf{q})$ denotes the mth feature functions. Typically, CRFs are able to handle a large set of feature-functions that captures a long span of contextual information. The details about the feature functions used in this work, including the syllabic features, will be described in Section 4. $Z(\mathbf{g}, \lambda)$ is the normalizing factor to ensure the conditional probabilities sum up to 1. Although higher order CRFs with more complex graphical model structures can be used for CRFs, it is more robust and computationally more efficient to train linear-chain CRFs. However, in order to apply linear-chain CRFs to G2P conversion, the letter-phoneme alignment needs to be provided explicitly. Therefore, it motivates us to look into combining CRFs and JMM to yield a hybrid system, which will be described in Section 3.

3. JMM-CRF Hybrid System

This section describes the proposed JMM-CRF hybrid system that integrates the JMM and CRF models in a probabilistic way so that the resulting system will benefit from the merits of both approaches. On one hand, JMM is able to analyze and decode multiple letter-phoneme alignments efficiently and captures a longer contextual information for the output labels. On the other hand, CRFs can easily incorporate complex feature functions and performs discriminative classification.

The JMM-CRF system works by first performing G2P conversion using the generative JMM model to obtain multiple hypotheses in the form of a lattice. Multiple letter-phoneme alignments are then extracted from the JMM lattice. For each alignment, a linear-chain CRF model is used to obtain a confusion network. A CRF lattice can then be constructed by taking the union of these confusion networks. Finally, the JMM and CRF lattices are combined together to form a final decoding lattice from which the best phoneme sequence is obtained. Since the lattices and confusion networks can be conveniently represented in the form of Weighted Finite-State Transducer (WFST), the manipulation of the JMM and CRF lattices can be easily achieved by means of WFST operations.

Fig. 1 shows a schematic diagram of the proposed JMM-CRF system. The JMM-CRF decoding involves three major steps: 1) Generation of the JMM lattices and extraction of multiple alignments; 2) Generation of the CRF confusion networks and the construction of the CRF lattices; and 3) Generation of the weighted combination of the JMM and CRF lattices and finding the best phone sequence. These steps will be described in the following sub-sections.

3.1. JMM Lattice Generation

We used the Sequitur G2P software provided by [5] to perform JMM decoding. Instead of generating the 1-best prediction, we generate multiple hypotheses in the form of a lattice, which encodes the most likely translations of the source sequence. Each path in the lattice uniquely corresponds to an alignment of the source sequence and a possible translation. The JMM lattices are represented as Weighted Finite-State Transducers (WFSTs), which are denoted by $\mathcal{F}_{\texttt{JMM}}$. The input and output labels of the arcs in $\mathcal{F}_{\texttt{JMM}}$ are given by the joint units. An example of $\mathcal{F}_{\texttt{JMM}}$ for the word 'BOX' is given as the first WFST in Fig. 1. From the JMM lattice, multiple letter-phoneme segmentations can be be extracted. In this case, there are two possible segmentations: B • O • X and B • O • _ • X. The underscore '_' is used to denote an empty symbol.

3.2. CRF Lattice Generation

Given a list of possible letter-phoneme segmentations, a linear-chain CRF is used to perform G2P for each segmentation. The open source CRF++[1] software was used for CRFs training and decoding. Instead of generating only the 1-best results, a CRF confusion network can be generated for each segmentation. There are multiple arcs spanning across each segment to represent the possible output phonemes. The arc weights are given by the CRF posterior probabilities. The second and third WFSTs in Fig. 1 show examples of CRF confusion networks obtained from two different segmentations. A CRF lattice, $\mathcal{F}_{\texttt{CRF}}$, can be obtained by performing a WFST union operation to all the CRF confusion networks.

[1] http://crfpp.sourceforge.net/

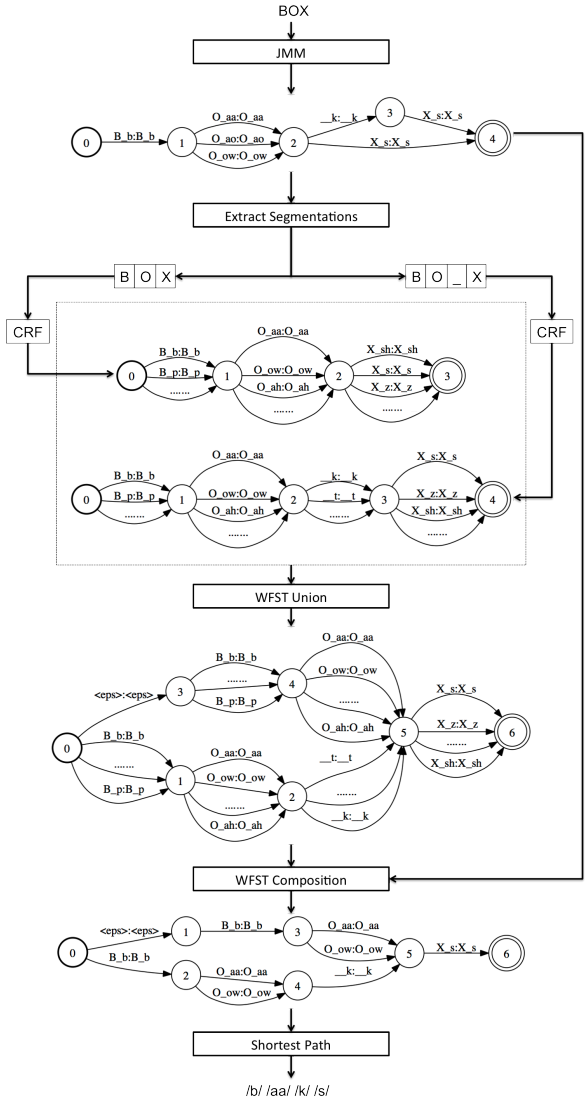

Figure 1: A schematic diagram of the JMM-CRF hybrid grapheme-to-phone conversion system. Not al arcs are shown and weights on WFST arcs are omitted for clarity.

3.3. A Weighted Combination of JMM and CRF Lattices

The JMM lattices ($\mathcal{F}_{\texttt{JMM}}$) and CRF lattices ($\mathcal{F}_{\texttt{CRF}}$) contain multiple hypotheses weighted by the JMM log likelihood scores and the CRF log posterior scores respectively. These two lattices can be easily combined together using the WFST composition operation. Such an operation essential finds the intersection of these two lattices and assign the weights as the sum of the JMM and CRF weights. Therefore, the proposed JMM-CRF system predicts the best phone sequence as follows:

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}\in Q} \{\alpha \log P(\mathbf{q}, \mathbf{g}|\theta_{\texttt{JMM}}) + (1-\alpha) \log P(\mathbf{q}|\mathbf{g}, \theta_{\texttt{CRF}})\}$$

where α is a weight parameter that can be adjusted to find a good compromise between the JMM and CRF models.

Table 1: Input features for /uw/ in the word 'CUBIC' with segmentation {C:k • _:y • U:uw • B:b • I:ih • C:k} and syllable structure {CU • BIC}. '_' denotes an *empty* symbol.

Orthographic features		Syllabic features	
Name	Example	Name	Example
1-gram	C, _, U, B, I, C	Syllable	CU
2-gram	C_, _U, UB, . . .	Beg_of_syl	0
3-gram	C_U, _UB, . . .	End_of_syl	1
4-gram	C_UB, _UBI, . . .	LetID_in_syl	2
1st order	y	SylID_in_wd	1

Table 2: *The Performance of the baseline JMM model and CRFs models with/without syllabic features.*

	CMUDict		CELEX	
Model	PER	WER	PER	WER
JMM	6.96	28.54	1.89	8.94
CRF_base	7.05	30.79	2.56	13.37
CRF_syl	6.55	29.84	2.10	11.09

4. Feature Functions for CRFs

This section describes the features used for CRFs, which are classified into two categories, as shown in Table 4. The *orthographic features* are the possible n-gram features extracted directly from the input letter sequence with a window length of 9 letters. Secondly, the *syllabic features* are a group of syllabification related information. In addition to using the syllable identity itself (referred to as **Syllable** in Tabel 4) as a feature, we also include two flag features, **Beg_of_syl** and **End_of_syl**, to indicate whether the current position is the beginning or end of a syllable. Besides, there are also two location-based features: 1) the letter index in the current syllable (**LetID_in_syl**) and the syllable index in the current word (**SylID_in_wd**). Table 4 gives an example of the input features associated with the output phone /uw/ at the third segment, which maps to the letter 'U'. It is located at the end of syllable 'CU' (**End_of_syl** = 1); it is also the second letter in the syllable (**LetID_in_syl** = 2) and the syllable 'CU' is the first syllable in the word (**SylID_in_wd** = 1).

5. Experiments

We evaluate the individual and hybrid systems on the CMUDict[2] and CELEX [14] English dictionaries. The CMUDict dictionary contains a total of 133252 words. After removing duplicate words, we randomly select 8000 words as development set and another 4000 words for evaluation. The remaining 111616 words are used as training data. The weight parameter, α, was tuned on the development set. On the other hand, the CELEX English dictionary contains 62821 unique words. In order to get reliable results on this relatively small lexicon, we employ a 10-fold cross-validation to evaluate the models. We first divide the data equally into 10 subsets, each time choosing one for development, one for evaluation and the remaining 8 subsets for training. We use the phone error rate (PER) and word error rate (WER) as evaluation metric.

We first build separate baseline JMM and CRFs models. The best JMM configuration is 9-gram for CMUDict and 8-gram for CELEX. The JMM system was used to generate the

[2]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

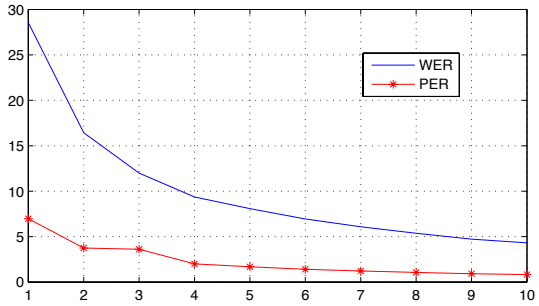

Figure 2: JMM Oracle Test on CMUDict. Horizontal axis indicates the depth of N-best list and vertical axis indicates the error rates.

Table 3: *Integrating performance on CMUDict evaluation set*

Model	PER	WER
JMM-CRF_base_10best	6.31	26.35
JMM-CRF_base_full	6.30	26.26
JMM-CRF_syl_10best	5.75	24.95
JMM-CRF_syl_full	**5.73**	**24.89**

alignment for both training and testing. The reference phone sequence was used for alignment during training. However, during decoding, the 1-best prediction from JMM was used instead. In order to compare the effectiveness of the syllabic features, two CRF models were constructed, with and without the syllabic features. Because there is no reference syllable for CMUDict, we get the syllabification information from [15]. They provide syllable boundary for phone sequence which is generated automatically using a structure SVM approach with 95% accuracy. We then map the phonetic syllable boundary to the orthographic syllable boundary according to the alignment. Experimental results of the baseline models on the evaluation sets are shown in Table 2. Both CRF_base (without syllabic features) and CRF_syl (with syllabic features) models performed slightly worse than JMM. Furthermore, adding the syllabic features to CRFs only gave marginal absolute WER reduction of 0.95% and 2.28% for CMUDict and CELEX respectively. Since we are going to combine JMM with CRFs, it is important to measure the quality of the JMM lattices in order to determine the appropriate lattice size for combination. In this work, we approximate the JMM lattices using N-best hypotheses. The oracle test results of different N-best lists for JMM on CMUDict are shown in Fig. 2. As N increases from 1 to 3, the PER and WER decrease from 6.96% and 28.54% to 2.61% and 11.99% respectively. When $N = 10$, the PER and WER reduced to 0.83% and 4.32% respectively.

Next, we evaluated the performance of the proposed JMM-CRF hybrid systems. We combine JMM with either CRF_base or CRF_syl. For each combination, we use either the 10-best hypotheses (JMM_10best) or the full lattice obtained from the JMM decoding (JMM_full). The WER and PER results are shown in Table 3. In general, the proposed JMM-CRF hybrid systems consistently outperformed the best individual systems. When combining using the full JMM lattices, the hybrid system achieved 2.28% and 3.65% absolute WER reduction without and with the syllabic features respectively. Using 10-best hypotheses instead of the full JMM lattices leads to only marginal

Table 4: *Integrating performance on CELEX evaluation sets*

	JMM-CRF_base		JMM-CRF_syl	
Set No.	PER	WER	PER	WER
1	1.65	8.09	1.60	7.74
2	1.84	8.68	1.69	8.15
3	1.81	8.47	1.57	7.53
4	1.72	8.42	1.68	8.1
5	1.69	7.98	1.49	7.18
6	1.71	8.31	1.50	7.45
7	1.92	8.74	1.65	7.9
8	1.65	8.18	1.58	7.88
9	1.76	8.58	1.55	7.86
10	1.57	7.50	1.39	6.89
Average	1.73	8.30	1.57	7.67

increase in error rates. This shows that keeping only the 10-best hypotheses from JMM is sufficient for subsequent combination with the CRF lattices. Therefore, in the following CELEX experiments, we only use the JMM 10-best outputs for combination. It is interesting to note that adding the syllabic features to the hybrid system gave a much larger absolute error rate reduction compared to the individual CRFs systems. The JMM-CRF_syl_full system gave the best results on the CMUDict evaluation, with 5.73% PER and 24.89% WER.

Finally, Table 4 gives the results of the proposed hybrid G2P conversion method for the individual cross-validation subsets. As before, we found that the hybrid system consistently outperform the individual systems. The JMM-CRF_base system achieved an average WER of 8.30%, which is 0.64% absolute lower than the baseline JMM. The WER further reduced to 7.67% when using syllabic features in the hybrid system. In general, JMM-CRF_syl consistently outperformed JMM-CRF_base both in terms of PER and WER across all cross-validation subsets.

6. Conclusions

In this work, we have proposed a JMM-CRF hybrid system for grapheme-to-phone conversion that combines the Joint Multigram Model (JMM) and the Conditional Random Field (CRF) classifier. JMM is an n-gram language model for the letter-phoneme joint units that can be used to predict the pronunciation given the letter sequence. In JMM-CRF, JMM is used to generate lattices that represent multiple letter-phoneme segmentations. A linear chain CRF model is used to rescore all possible segmentations in the JMM lattice. Both the JMM and CRF scores are then used to obtain the best hypotheses from these lattices. The outputs from JMM and CRF can be conveniently represented in the form of Weighted Finite State Transducers (WFSTs). Composing the two WFSTs and finding the shortest path leads to the desired solution. Furthermore, syllabic information can be easily incorporated into CRFs as additional features. Experimental results on the CMUDict and CELEX databases show that the proposed JMM-CRF hybrid system consistently outperformed both the individual systems. The syllabic features have also been experimentally shown to improve G2P performance, especially when integrating with JMM

7. References

[1] T. Sejnowski and C. Rosenberg, "Parallel networks that learn to pronounce english text," *Complex systems*, vol. 1, no. 1, pp. 145–168, 1987.

[2] H.-U. Hain, "Automation of the training procedure for neural networks performing multilingual grapheme to phoneme conversion," in *Proceedings Eurospeech*, 1999, pp. 2087–2090.

[3] J. Lucassen and R. Mercer, "An information theoretic approach to the automatic determination of phonemic baseforms," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 9. IEEE, 1984, pp. 304–307.

[4] P. Taylor, "Hidden markov models for grapheme to phoneme conversion," in *Ninth European Conference on Speech Communication and Technology*, 2005.

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

[6] D. Wang and S. King, "Letter-to-sound pronunciation prediction using conditional random fields," *Signal Processing Letters*, vol. 18, no. 2, pp. 122–125, 2011.

[7] S. Hahn, P. Lehnen, and H. Ney, "Powerful extensions to crfs for grapheme to phoneme conversion," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*. IEEE, 2011, pp. 4912–4915.

[8] I. Illina, D. Fohr, and D. Jouvet, "Grapheme-to-phoneme conversion using conditional random fields," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[9] D. Jouvet, D. Fohr, and I. Illina, "Evaluating grapheme-to-phoneme converters in automatic speech recognition context," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*. IEEE, 2012.

[10] V. Demberg, H. Schmid, and G. Mohler, "Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion," in *Annual Meeting-association for Computational Linguistics*, vol. 45, no. 1, 2007, p. 96.

[11] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[12] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 1. Association for Computational Linguistics, 2003, pp. 134–141.

[13] X. Wang, L. Xie, B. Ma, E. Chng, and H. Li, "Modeling broadcast news prosody using conditional random fields for story segmentation," *Proceedings of APSIPA ASC*, pp. 253–256, 2010.

[14] R. Baayen, R. Piepenbrock, and L. Gulikers, "CELEX2," *Linguistic Data Consortium*, 1996.

[15] S. Bartlett, G. Kondrak, and C. Cherry, "On the syllabification of phonemes," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 308–316. [Online]. Available: http://www.aclweb.org/anthology/N/N09/N09-1035