Adaptive Statistical Utterance Phonetization for French

Gwénolé Lecorvé, Damien Lolive

ADAPTIVE STATISTICAL UTTERANCE PHONETIZATION FOR FRENCH

Gwénolé Lecorvé, Damien Lolive

gwenole.lecorve@irisa.fr, damien.lolive@irisa.fr
IRISA/Université de Rennes 1, Lannion, France

ABSTRACT

Traditional utterance phonetization methods concatenate pronunciations of uncontextualized constituent words. This approach is too weak for some languages, like French, where transitions between words imply pronunciation modifications. Moreover, it makes it difficult to consider global pronunciation strategies, for instance to model a specific speaker or a specific accent. To overcome these problems, this paper presents a new original phonetization approach for French to generate pronunciation variants of utterances. This approach offers a statistical and highly adaptive framework by relying on conditional random fields and weighted finite state transducers. The approach is evaluated on a corpus of isolated words and a corpus of spoken utterances.

***Index Terms*—** Utterance phonetization, pronunciation variant modelling, phoneme lattices, conditional random fields, weighted finite state transducers.

1. INTRODUCTION

Pronunciation generation aims at predicting a sequence of phonemes based on a graphemic input. Most of the time, this task is limited to grapheme-to-phoneme (G2P) conversion of isolated words. Phonetizing sentences thus usually consists in concatenating pronunciations of their constituent words. However, this approach is too weak for some languages, like French, where transitions between words imply pronunciation modifications. Moreover, this approach makes it difficult to consider global pronunciation strategies, for instance to adapt a speech processing system to a specific speaker or accent. This objective is all the more important since these adaptation tasks have become a major problem, especially in the field of text-to-speech synthesis (TTS) [1].

To overcome these problems, this paper presents a new phonetizer for French. This phonetizer brings three main contributions. First, it introduces the notion of elision model to model intra-word variants. Second, it integrates phonological contexts to model inter-word variants. Finally, it is able to generate probabilistic phoneme lattices from sentences, and not only from isolated words. To do so, the phonetizer relies on conditional random fields (CRFs) to estimate phoneme probabilities of isolated words and on weighted finite state transducers (WFSTs) to allow transitions between words. Results are phoneme lattices from which phonetizations can be derived.

The potential of this phonetizer is very high. Generated phoneme lattices offer a lot of flexibility since transitions can be rescored using various pronunciation models, e.g., to perform speaking style or accent adaptation. Nonetheless, this paper focuses on the sole presentation of the phonetizer and exhibits first results without adaptation. The latter is kept for future work. More generally, it is important to highlight that this paper does not seek to outperform state-of-the-art models and tools. Rather, the main goal is to define the framework in a generic way and to demonstrate the method on French. Still, this framework can be easily extended and completed with additional models. Moreover, it does not rely on expert rules and can thus be ported to other languages with only minor knowledge. Finally, the proposed framework can also tolerate uncertainty in the input utterance, for instance to handle multiple tokenizations.

This paper is organized as follows: Section 2 first draws an overview of the domain; Section 3 introduces our G2P conversion method for isolated words before extending it to utterances in Section 4. Experiments are presented for isolated words and utterances on the pronunciation lexicon MHATLex and on a speech corpus.

2. STATE OF THE ART

Phonetization generation has been widely studied for a long time, particularly in automatic speech recognition (ASR) and in TTS. Most ASR and TTS systems mainly rely on hand-crafted pronunciation lexicons for common words and fall back on automatic G2P converters for out-of-vocabulary words (OOVs), i.e. words which are not in the lexicon. Many strategies have been proposed for G2P conversion in the literature: rule-based methods [2, 3], statistical approaches [4, 5], and other varied techniques [6, 7]. Among those, statistical approaches have recently shown very interesting performance while also providing advantages of statistical frameworks, especially the possibility to interpret and adapt scores of the generated pronunciations. Two main methods are competing. On the one hand, joint multigrams methods rely on sequences of grapheme-phoneme pairs[1] whose probabilities are usually obtained using language modelling methods [4, 8]. A comparative overview of these methods can be found in [9]. On the other hand, CRFs have proven to efficiently address the G2P problem [5, 10, 11, 12]. They can now be considered as the state of the art.

The pronunciation of utterances, i.e. word sequences, has been more seldomly studied. In ASR, the introduction of WFSTs as a mean to decode speech signals has come along with a new representation of alternative pronunciations of words [13]. Especially, [14] proposed to represent utterances, pronunciations and possible variations on them as WFSTs which can be composed and searched to extract pronunciation variants. In isolated word G2P methods, WFSTs and phoneme lattices are also used to represent phonetization alternatives [15, 16] or even directly CRFs [17]. The philosophy of the current paper is very close as it combines CRFs and WFSTs. The work presented here is however different from [14] since OOVs and phoneme elisions are introduced here. Furthermore, [14] focuses on English whereas our work is achieved on French, which is phonologically different.

The phonetization of isolated words is presented in the next section before moving to utterances in Section 4.

[1] Several graphemes and several phonemes can be considered in one pair.

3. G2P CONVERSION OF ISOLATED WORDS

G2P conversion of isolated words consists in predicting a sequence of phonemes based on a given grapheme sequence. It can be seen as a classical supervised classification problem where labelled training data is required and a type of model has to be defined. In this work, training data is generated using automatic grapheme-phoneme alignments of a pronunciation dictionary and CRFs are chosen for modelling. This section presents the alignment strategy, the model training and finally exhibits some results on isolated words.

3.1. Grapheme-phoneme alignment

In the G2P conversion problem, each grapheme must be aligned with phonemes in the training data. While some related work relies on hidden Markov models to perform this step [5], a many-to-many alignment algorithm has been used in our work[2]. No sensible difference in performance is known between the two approaches. Many-to-many (M2M) alignment has been chosen for practical reasons since it is judged as more flexible. M2M alignment seeks to maximize a probabilistic function of graphemic and phonetic tuples being aligned [18]. Major parameters to be tuned are the probabilistic function itself and the maximal size of tuples. Various combinations have been tested in a preliminary work and results showed that the best objective function is the joint probability of graphemes and phonemes and that tuples of maximal size 2 leads to the lowest error rates. This setting is used for all the experiments in this paper.

3.2. Conditional random field training

CRFs are based on exponential models from which conditional probabilities of a target sequence Y given a source sequence X can be derived. Using these probabilities, a decoding algorithm produces the most probable sequence Y^*. The underlying exponential model is defined as follows:

$$\Pr(Y|X) = \frac{1}{Z(X,Y)} \exp \sum_{i \in I} \lambda_i F_i(X,Y) \,, \qquad (1)$$

where $Z(X,Y)$ is a normalization factor, I is the set of considered features for the problem, and λ_i coefficients are weights which are optimized during training. Finally, F_i functions are feature functions that return 1 if a given property in X and Y is observed, 0 otherwise. Feature functions carry one of the strengths of CRFs since they allow for a wide range of parameters to characterize a problem.

Given alignments between graphemes and phonemes, a G2P CRF is trained by using as training features the sole dependencies between each phoneme and its corresponding grapheme or grapheme n-grams. In the literature, using grapheme n-grams instead of single graphemes had led to satisfactory results in English [10] and French [5]. However, as French contains many homographs with different pronunciations[3], additional information is required, typically part-of-speech. [5] has shown that this feature can be limited to differentiating verbs from other morphosyntactic classes.

In the remainder, we denote as $\mathbf{g}$ the grapheme n-gram used to predict a given phoneme and as $\mathbf{o}$ the other features derived from a word. After training, the G2P CRF is used to return the best or the few best phonetization hypotheses. Additionally, the posterior probability $\phi(p|\mathbf{g},\mathbf{o})$ can be obtained for each phoneme p.

[2] `https://code.google.com/p/m2m-aligner/`.

[3] A typical example is the written form "*président*", either pronounced /pʁezidɑ̃/ if it is a noun or /pʁezid/ if this is the inflected form of the verb *présider*.

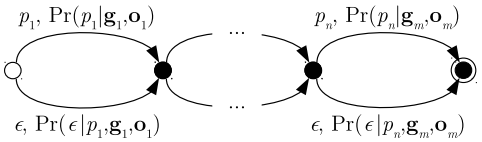

Fig. 1. Phoneme lattice of an isolated word pronunciation.

3.3. Elision modelling

As in other languages, some phonemes in French can be elided, i.e. omitted. These elisions depend on various information such as the phonological context, the type of speech, grammatical rules or exceptions, etc. The most representative phenomenon to illustrate this variability is the phoneme schwa (/ə/) which can be elided almost all the time, but not always. For instance, the word *semaine* (week) can be pronounced /səmɛn/ or /smɛn/. Notice that both pronunciations do not carry the same expressiveness as the short one is rather uttered in an informal communication context. Additionally, the last grapheme *e* can also be pronounced /ə/ when followed by a consonantic sound. Hence, the sentence "*la semaine finit*" can be pronounced /lasəmɛnfini/ or /lasəmɛnəfini/. In some types of speech, pronouncing such final schwas is a rule, e.g., when reciting poetry. Still, not all schwas are optional. For instance, the word *Bretagne* (Brittany) is compulsorily pronounced /bʁətaɲ/ as omitting the schwa would lead to uttering three incompatible consonants. Similar phenomena exist for other phonemes, especially the so-called *liaisons* when considering word linkages.

In this paper, we propose to train another CRF, referred to as elision CRF, to predict phoneme elisions. During training, for each phoneme in a word pronunciation, the elision label to be learned is either set to "opt" if the phoneme is optional, i.e. it is possibly pronounced *or* not pronounced, or set to "mand" if it is mandatory. In addition to the graphemic features $\mathbf{g}$ and other features $\mathbf{o}$ used to train the G2P CRF, phoneme n-grams are also used here. After training, the elision probability $\varepsilon(p,\mathbf{g},\mathbf{o})$ of a given test phoneme p is derived from the elision CRF as follows:

$$\varepsilon(p,\mathbf{g},\mathbf{o}) = \begin{cases} 0.5 \times Q_{\mathrm{E}}(p,\mathbf{g},\mathbf{o}) & \text{if } C_{\mathrm{E}}(p,\mathbf{g},\mathbf{o}) = \text{opt}, \\ 1 - Q_{\mathrm{E}}(p,\mathbf{g},\mathbf{o}) & \text{if } C_{\mathrm{E}}(p,\mathbf{g},\mathbf{o}) = \text{mand} \,, \end{cases} \qquad (2)$$

where $C_{\mathrm{E}}(p,\mathbf{g},\mathbf{o})$ denotes the label returned by the elision CRF for p and $Q_{\mathrm{E}}(p,\mathbf{g},\mathbf{o})$ is the posterior probability of this label. Since only 2 labels are possible, $Q_{\mathrm{E}}(p,\mathbf{g},\mathbf{o})$ ranges in $[0.5,1]$ for any returned, i.e. highest probability, label. According to Eq. 2, $\varepsilon(p,\mathbf{g},\mathbf{o})$ thus ranges in $[0,0.5]$. This is an empirical choice to avoid the elision model to completely erase decisions made by the G2P model. This choice can still be modified if *a priori* knowledge about the phonetization strategy is given.

As a consequence, the probability of a phoneme can be reformulated as:

$$\Pr(p|\mathbf{g},\mathbf{o}) = \phi(p|\mathbf{g},\mathbf{o}) \times (1 - \varepsilon(p,\mathbf{g},\mathbf{o})) \,, \qquad (3)$$

and the complementary probability of skipping p is:

$$\Pr(\epsilon|p,\mathbf{g},\mathbf{o}) = \phi(p|\mathbf{g},\mathbf{o}) \times \varepsilon(p,\mathbf{g},\mathbf{o}) \,, \qquad (4)$$

where ϵ refers to "no phoneme". Using these probabilities, a phoneme lattice can be built for each given phoneme sequence and the path with the highest probability is chosen as the best phonetization. The architecture of such a lattice is drawn in Fig. 1. Edges are labelled with a phoneme or with ϵ, and weighted with the probabilities from Eq. 3 and 4, respectively. This principle can be

extended with n-best lists instead of the sole 1-best hypothesis returned by the G2P model. After applying the elision model on each hypothesis, a new lattice can be built as the union of all alternative phoneme sequence hypotheses.

3.4. Experiments on isolated words

The G2P conversion method has been applied on the MHATLex corpus [19]. This corpus lists $450,000$ words along with a total of $710,000$ pronunciations[4]. Each word is given with its POS and each pronunciation includes elision possibilities and the phonological contexts for which the pronunciation stands. This corpus is a more detailed version of the BDLex corpus, used in [5]. The phonological contexts are disregarded in this first series of experiments. They will be used in Section 4. The corpus has been partitioned into a training set (75 %), a development set (5 %), and a test set (20 %). The $2,000$ most frequent words in French have been to put into the training set as these words will never be OOVs in real life applications and they also tend to have irregular pronunciations. Moreover, words sharing the same lemma[5] have been gathered into the same set to avoid the sets to be morphologically too similar. G2P and elision CRFs are learned on the training set using the development set to define the learning stopping criterion while evaluations are carried out on the test set. The CRF training toolkit is Wapiti[6] [20].

Different feature sets have been tested for G2P training. Their components are grapheme n-grams (grapheme g_i surrounded by a window of $\pm N$ graphemes) and the verb/non-verb (POS) information. Different window sizes have been tested while POS is always used. In addition to these features, elision models take into account the current phoneme p_j and its $\pm N$ surrounding phonemes[7]. Table 1 reports the phoneme error rates (PERs) and word error rates (WERs) on the test set for various window sizes and with or without the elision model. Results are compared to those of Liaphon, which is the most widely used utterance phonetization system for French [2]. Liaphon relies on thousands of hand-crafted rules covering general pronunciation phenomena as well as exceptions. Liaphon's version used in the experiments is optimized for TTS. On the contrary, results from the CRF-based approach in [5] are not exposed because the corpus and the partitioning strategy are partially different.

First, it appears from the results that our G2P CRFs tend to achieve similar results as Liaphon, though they are slightly worse. Increasing the grapheme window size brings improvements. However, after size 2, it appeared in our experiments that the quality CRFs was degraded. We think that this is because the training set contains many nearly similar words due to the constraint on lemmas, which leads to overtraining. Second, the use of elision models brings variability in the phoneme lattices without significantly altering nor improving the G2P results.

4. PHONETIZATION OF UTTERANCES

Standard phonetization tools are focused on isolated words and consider utterances as a concatenation of such words. However, in many languages, word transitions introduce phonological modifications.

[4]Hence, many words are provided with several pronunciation variants.

[5]A lemma is a canonical form of a word. For example, plural nouns are reduced to their singular form, conjugated verbs are reduced to their infinitive form...

[6]`http://wapiti.limsi.fr/`

[7]N is set to the same value as the grapheme window size to avoid introducing another parameter.

Input features	PER (%)	WER (%)
Grapheme (no window) + POS	5.8	29.9
+ elision model	5.7	29.5
Graphemes (± 1) + POS	2.6	11.3
+ elision model	2.4	11.6
Graphemes (± 2) + POS	1.8	9.0
+ elision model	1.9	9.3
Liaphon	1.3	6.8

Table 1. PERs and WERs on the test set of MHATLex.

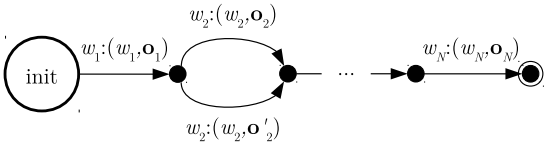

Fig. 2. WFST of an utterance. All edges are weighted with probability 1.

In French, liaisons are the most patent phenomenon and are usually handled using expert rules. In this section, we propose (i) to fully model word transitions by introducing phonological contexts in the previously proposed statistical framework and (ii) to represent sentences as a WFST in order to keep track of all possible pronunciation variants. After presenting how the G2P and elision CRFs can integrate phonological contexts and how utterances can be modeled using WFSTs, experiments on a speech corpus are reported.

4.1. Introduction of phonological contexts

A given word w_i influences the pronunciation of its preceding and succeeding words w_{i-1} and w_{i+1}. Reciprocally, the pronunciation of w_i depends on those of w_{i-1} and w_{i+1}. The influencing characteristics of these neighbouring words is referred to as the phonological context. Let us denote as l_i the information transmitted on the left by w_i to w_{i-1}, and r_i the information transmitted on the right to w_{i+1}. In a symetric manner, the pronunciation of w_i depends on r_{i-1} and l_{i+1}. Hence, we propose to integrate r_{i-1} and l_{i+1} as new features in the training process of the G2P and elision CRFs.

4.2. WFST representation

The basic idea in producing a phoneme lattice for a given utterance consists in building a WFST representation of the utterance and composing it with a WFST representation of its word pronunciations. Given an utterance of N words, the utterance WFST representation simply consists in a linear chain of transitions where each word w_i is transduced into its parametrization $(w_i, \mathbf{o}_i)$ (Fig. 2). Potentially, a given word may lead to different parametrizations. In this case, which is illustrated with the word w_2, alternative paths in the WFST can be built. By default, all transition probabilities are set to 1.

Building the lexicon WFST is more complex as word transitions have to be carefully modeled. Fig. 3 illustrates the architecture of such a WFST. Considering each parametrized word $(w_i, \mathbf{o}_i)$ to be phonetized, several phonetizations can be relevant according to the phonological context where the word will take place. These phonological contexts are represented as nodes (a_i, b_j) from which and to which each phonetization is linked. Between these nodes, following the principle of Fig. 1, each phoneme sequence is represented as a chain where $(w_i, \mathbf{o}_i)$ is consumed by the first edge and phonemes $p_{i,j}$ are output in the remaining

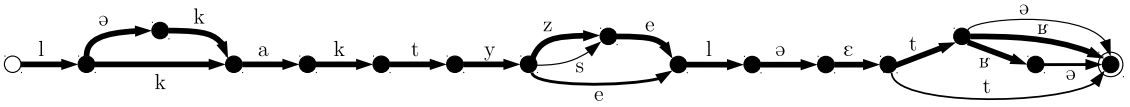

Fig. 4. Phoneme lattice generated by the phonetizer for the sample utterance "*le cactus et le hêtre*" ("the cactus and the beech"). For clarity, probabilities are represented with different line sicknesses. Boldest lines correspond to probability 1.

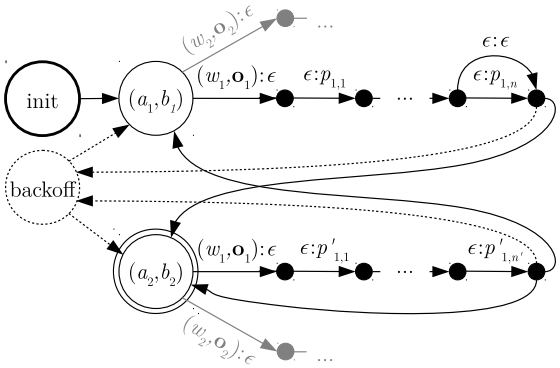

Fig. 3. WFST of the lexicon. Unlabelled edges are ϵ:ϵ transitions. Probabilities are not drawn for clarity.

edges. Elisions are represented as ϵ-transitions. Finally, word transitions are handled as follows: each contextualized pronunciation $(r_{i-1}, w_i, \mathbf{o}_i, p_{i,1}, \ldots, p_{i,n}, l_{i+1})$ is linked to all possible context nodes (a, r_{i-1}) and (l_{i+1}, b), where a and b spans the contexts l_i and r_i which can be transmitted by w_i to the left and to the right, respectively. All pronunciations are also linked to a backoff node to allow irregular words transitions. Edges to context nodes are weighted with probability 1 while those to the backoff node are weighted with a penalty empirically set to $\exp(-10)$. Finally, according to their meaning, some context nodes are final states.

All the pronunciations of a given utterance can thus be stored into one single WFST. Pronunciations are either derived from the pronunciation dictionary or from the contextualized word phonetizer presented in Section 4.1 for OOVs. The probability of in-vocabulary word pronunciations are set to 1 or 0.5 whether the phoneme is mandatory or optional and a complementary epsilon transition is also built with the same probability. By composing the utterance WFST with the lexicon WFST, a phoneme lattice is obtained and is decoded to generate the best or the few best utterance pronunciations.

4.3. Experiments on utterances

Left phonological contexts have been derived from information provided in the MHATLex corpus. Two contexts r_{i-1} are considered: either the previous word ends with an open syllable or with a closed syllable. The options for contexts l_{i+1} are more varied: the next word starts either with a consonant, a non-consonant (semi-vowel or vowel), a nasal or non-nasal phoneme, or it prevents from liaisons or there is no next word, i.e. end of sentence. This last context is the only one enabling a context node to be a final state. G2P and elision CRFs have been retrained on the training set augmented with context information.

The proposed approach has been applied on a speech corpus of about $1,400$ utterances for a total of 12K words. This corpus comes with a manually checked phonetization of each utterance. Utterances have been phonetized using the best CRF models from Section 3. Four configurations have been tested: the use of graphemes and POS tags alone, graphemes and POS with elision or phonological contexts separately, and all information together. Results are measured in terms of PERs again but WERs are replaced by sentence error rates (SERs) as the objective is to get a fully correct phonetization of each utterance. Results are presented in Table 2 and compared to those of Liaphon on the same corpus. First, PERs are much higher than on isolated words. This clearly shows the difficulty to model utterance pronunciations. Then, we can see from the various configurations that, separately, the elision model and the phonological contexts bring significant improvements over the baseline and their combination leads to further improvements. Finally, the results from our technique are worse than those of Liaphon. We think that this is logical because many alternative paths with equal probabilities are stored in phoneme lattices. These competing paths come from the in-vocabulary words for which elisions and other variants are all considered with the same weight, which is not true in the language. On the contrary, Liaphon includes assumptions about elisions and liaisons. An example of a (pruned) phoneme lattice is shown in Fig. 4 for the utterance "*le cactus et le hêtre*", where the words *cactus* and *hêtre* are OOVs. Alternative paths are clearly appearing. Better results could probably be achieved by rescoring phoneme lattices, e.g., with a phoneme-based language model trained on an excerpt of the speech corpus. This will be done in the very near future. Still, we demonstrate with this example the potential of our approach.

Input features and models	PER (%)	SER (%)
Graphemes (± 2) + POS	22.6	88.4
+ elision model (no phonological context)	16.8	89.2
+ phonological contexts (no elision model)	17.7	85.6
+ elision model + phonological contexts	16.4	87.7
Liaphon	13.2	57.4

Table 2. PERs and SERs on the speech corpus.

5. CONCLUSION AND PERSPECTIVES

This paper presented a new utterance phonetization method for French. The main goal of this original method is to produce phoneme lattices which can be easily post-processed to fit specific requirements, like a given speaking style or accent, especially for TTS. The method relies on the use of CRF models to phonetize isolated words, elide phonemes and take into account phonological contexts, and on WFSTs to extend the phonetization to utterances.

Many perspectives are open thanks to the proposed method. First, utterance WFSTs can be made more complex by considering several tokenizations. This is for instance useful to model word contractions or acronyms. Second, post-processings of the produced phoneme lattices should be improve many TTS applications where a specific expressiveness or speaking style is needed, e.g., video games, audiobooks, language learning software. Finally, the use of graphs could be pushed further by integrating the phoneme lattices directly into TTS engines in order to offer them more choices.

6. REFERENCES

[1] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer, 2008.

[2] F. Béchet, "LIA_PHON : un système complet de phonétisation de textes," *Traitement Automatique des Langues (TAL)*, vol. 42, no. 1, pp. 47–67, 2001.

[3] V. Claveau, "Letter-to-phoneme conversion by inference of rewriting rules," in *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 1299–1302.

[4] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, 2008.

[5] I. Illina, D. Fohr, and D. Jouvet, "Grapheme-to-Phoneme Conversion using Conditional Random Fields," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 2313–2316.

[6] J. R. Bellegarda, "Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy," *Speech Communication*, vol. 46, no. 2, pp. 140–152, 2005.

[7] A. Laurent, P. Delglise, and S. Meignier, "Grapheme to phoneme conversion using an SMT system," in *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 708–711.

[8] J. R. Novak, N. Minematsu, and K. Hirose, "WFST-based grapheme-to-phoneme conversion: open source tools for alignment, model-building and decoding," in *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, Donostia-San Sebastián, Spain, 2012, pp. 45–49.

[9] S. Hahn, P. Vozila, and M. Bisani, "Comparison of grapheme-to-phoneme methods on large pronunciation dictionaries and LVCSR tasks," in *Proceedings of Interspeech*, Portland, OR, USA, 2012.

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

[11] S. Hahn, P. Lehnen, and H. Ney, "Powerful extensions to CRFs for grapheme to phoneme conversion," in *Proceedings of ICASSP*, Prague, Czech Republic, 2011, pp. 4912–4915.

[12] P. Lehnen, S. Hahn, V.-A. Guta, and H. Ney, "Hidden conditional random fields with M-to-N alignments for grapheme-to-phoneme conversion," in *Proceedings of Interspeech*, Portland, OR, USA, 2012, pp. 2554–2557.

[13] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," in *Proceedings of the Intl Workshop on Automatic Speech Recognition : Challenges for the Next Millenium*, 2000.

[14] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," *Speech Communication*, vol. 46, no. 2, pp. 189–203, 2005.

[15] N. Bodenstab and M. Fanty, "Multi-pass pronunciation adaptation," in *Proceedings of ICASSP*, vol. 4, 2007, pp. 865–868.

[16] T. Polyákova and A. Bonafonte, "Introducing nativization to spanish TTS systems," *Speech Communication*, vol. 53, no. 8, pp. 1026–1041, 2011.

[17] P. Lehnen, S. Hahn, and H. Ney, "N-grams for conditional random fields or a failure-transition (φ) posterior for acyclic FSTs," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 1437–1440.

[18] S. Jiampojamarn, G. Kondrak, and T. Sherif, "Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion," in *Proceedings of HLT-NAACL*, Rochester, NY, USA, 2007, pp. 372–379.

[19] G. Pérennou and M. De Calmes, "MHATLex: Lexical resources for modelling the French pronunciation," in *Proceedings of LREC*, Athens, Greece, 2000.

[20] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 504–513.