APPLYING PRONUNCIATION MODELING TECHNIQUES TO FRENCH

Daniel Gildea and Eric Fosler-Lussier

International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
University of California at Berkeley, EECS Department, Berkeley, CA 94720
Tel: (510) 643-9153, FAX: (510) 643-7684, Email: {gildea, fosler}@icsi.berkeley.edu

ABSTRACT

This paper describes the use of data-driven pronunciation modeling techniques to substantially reduce error in a large vocabulary hybrid ANN/HMM speech recognition system for French. Pronunciations from canonical transcriptions are combined with pronunciations from a text-to-speech system. Forced alignments are used to calculate pronunciations' probabilities and to eliminate bad pronunciations. Phonetic rules are used to smooth probabilities and ensure the inclusion of pronunciations that may not be present in the training data. Issues concerning the choice of phonemes for the system are also discussed.

1. INTRODUCTION

In an automatic speech recognition system, words are usually represented by a set of acoustically based sub-word units, often phones or triphone. Much effort from the ASR community has gone into determining pronunciation dictionaries that provide this word to acoustic unit mapping. Several methods are commonly in use; each of them has advantages and disadvantages.

One primary source of pronunciations is a hand-built lexicon, where experts carefully craft pronunciations for each word [6]. When care is taken to minimize dictionary confusions and ensure consistency of pronunciation across similar words, the resulting lexicon is often excellent. However, the work, being manual in nature, is very time consuming; incorporating a new vocabulary into a recognizer can often be prohibitively expensive.

The difficulty of obtaining accurate transcriptions for all words in a large vocabulary system, as well as the variability of natural speech, has lead to several techniques for automatically deriving dictionary pronunciations from data. A straightforward approach is to generate pronunciations with a phone recognizer for use in the word recognizer, pruning based on confidence or counting thresholds ([10] *inter alia*). This has the advantage of learning pronunciations for frequent words well, but it doesn't learn phonological correlations across words. One can also generate pronunciations using phonological rules [11, 4], and learn probabilities for words based on rule probabilities estimated from the corpus. This has the opposite effect: specific pronunciations for words are not learned, but rather generalities across a corpus.[1]

Other lexicon-building techniques are also used in the community; decision trees and finite state transducers learned from phone recognition can give a "smoothed" phonetic representation ([9] *inter alia*). Text-to-speech systems are sometimes used to find pronunciations for out of vocabulary words (*e.g.* [12]). In summary, there are many ways to generate pronunciations for ASR systems.

In this work, we combine some of these methods into one pronunciation learning system. This work continues the hybrid ANN/HMM system development for French presented in [2]. We present a single system which learns from the pronunciations present in training data, and uses phonetic rules[2] to generalize to words not seen in the training data, extending the work of Tajchman *et al.* [11]. The question of choosing pronunciations also led us to investigate various phonemic representations for French. We show that techniques developed for pronunciation modeling in a hybrid ANN/HMM system for English can also be successfully applied to French, despite significant differences in the phonology of the two languages.

2. METHODOLOGY

The dataflow of the pronunciation modeling system is shown in Figure 1. Given a training set and an initial set of pronunciations, a forced alignment is performed, allowing each instance of a word to use the pronunciation in the dictionary that best fits the acoustics. Probabilities for each word's various pronunciations are re-estimated by counting the frequency with which each pronunciation occurs in the data. Another forced alignment is performed in order to retrain the neural net phone probability estimator. This process can be iterated several times, alternating re-estimation of the lexicon and retraining of the net. In practice, one or two iterations are generally sufficient to reach convergence.

In addition to simply calculating probabilities for each pronunciation, probabilities are also calculated for phonetic rules. These rules, which are written by hand, are applied to baseform pronunciations by the system to generate possible pronunciation variants for each word. Each form in the dictionary is tagged to indicate whether, in its derivation, each optional rule whose conditions were met was applied or not. By counting the tags on the forms chosen by the forced alignment, an overall probability for each rule's application is calculated. These rule probabilities are then used to calculate a rule-based probability for each rule-derived pronunciation. The probability of each rule which applies or fails to apply in a given pronunciation is multiplied, and the values for all pronunciations of a given word are then normalized to sum to one.

For words that do not appear in the training data, the rule-based pronunciation probabilities are used directly. For words which oc-

[1] One can compensate for this to some extent by learning word-specific rules.

[2] We use term phonetic rather than phonological rules because it more accurately describes the rules used in the present work. There is no difference for the functioning of the system.

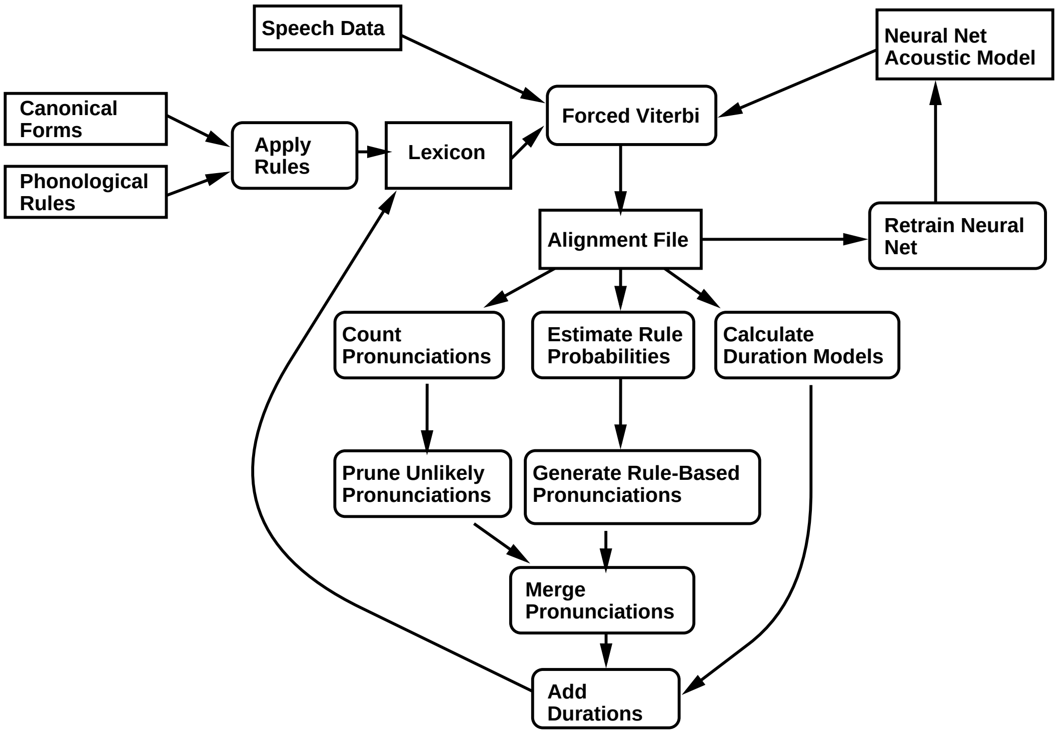

Figure 1: Dataflow of the Pronunciation Modeling System

cur rarely in the training data, we may want to include rule-based pronunciations in order not to leave out variant that may simply not be found in the training set. For common words, the empirically observed pronunciations should be given more credence in order to capture any effects peculiar to the word in question — in particular pronunciation variants of common function words. We accounted for this by weighting the rule-based pronunciations as the equivalent of 10 observed pronunciations. Thus, for a word seen 10 times in the training data, the final probabilities were the mean of the rule-based and empirical probabilities, whereas for a word seen 90 times in the training data, the final probabilities were a weighted mean with a weight of 1/10 for the rule-based and 9/10 for the empirical probabilities.

Context-dependent durations for each phoneme were also recalculated from the forced alignment. An average duration for each phoneme given the phonemes to the left and right is calculated. This average duration is then used to calculate the number of states assigned to the phoneme in the HMM pronunciation model.

Pronunciations that have been assigned low probabilities by the process described above may be pruned — that is, eliminated from the dictionary entirely. A number of different pruning methodologies are possible; we choose to prune a fixed proportion of the probability mass for each word. Based on an examination of our dictionaries, we set this constant at 10%. Thus the least likely pronunciations are eliminated until a pronunciations adding up to a total probability of 10% have been removed. This helps to filter out errorful pronunciations, and reduces confusibility between vocabulary items. Pruning was applied to pronunciations found in the training data, before they were combined with rule-derived pronunciations. This is because the pronunciations generated by applying rules to canonical forms were considered to be more reliable.

3. RESULTS

Experiments were performed using the BREF corpus, which contains studio-quality read speech from the newspaper Le Monde. The system used was a hybrid ANN/HMM system [1] with RASTA-PLP features[5]. Twelfth order features were used, along with their deltas and delta energy and delta-delta energy for a total of 26 features. Features were calculated every 10ms over a 25ms window. The neural net examined a window of nine frames, for a total of 234 inputs, and had 1000 hidden units. A set of 36 phonemes was used, including one silence phoneme. The NOWAY [8] decoder was used, and a lexicon of the 66,691 words was chosen by supplementing the 65,000 most common words in French with the words necessary to ensure no out-of-vocabulary words in the training or test data. The training data were 3737 utterances with a total of 69,997 words and 7.5 hours of speech. The test data were 242 utterances with a total of 3662 words and 20 minutes of speech. A trigram grammar was trained on roughly 80 million words from *Le Monde* over the years 1990-1992 (the month in which the BREF recordings were made was excluded).

A major advantage of a data-driven pronunciation modeling system is that it allows the user to include pronunciations from a variety of sources and let the system sort out which are good and which are bad on its own. We combined pronunciations from two main sources: a dictionary of canonical pronunciations written by experts, and a text-to-speech system. The dictionary used was BDLex (distributed by the European Language Resources Association), which has pronunciations for 50,000 French stems in a total of over 300,000 inflected forms. The dictionary includes information about optional liaison consonants and the optional “e muet” sound. Suplementing BDLex was essential, as BDLex contains no proper nouns. Two types of TTS pronunciations were used, which we call in-context and out-of-context pronunciations. A dictionary of out-of-context pronunciations was generated by presenting the TTS [3] with all of the words not found in BDLex, in isolation and lower case. The TTS was also used to generate an in-context pronunciation of each complete utterance in the training data. These pronunciations were generally more accurate than the out-of-context pronunciations because the TTS was able to use context words, punctuation, and capitalization. These full utterance pronunciations were broken into pronunciations for each component word by aligning the utterance pronunciation with dictionary pronunciations.

Two simple phonetic rules were used: one to estimate the probability with which a liaison consonant at the end of a word is produced, and one to estimate the probability with which the optional schwa of the French "e muet" is produced. Both of these rules are simplifications of the underlying phonological processes. Fairly complex syntactic phenomena govern environments in which liaison is required, optional, or prohibited before a following word-initial vowel in French. Our system is not sensitive to these constraints, or even to the fact that liaison can only occur before a word starting with a vowel. It simply calculates an overall probability of any liaison consonant being produced, regardless of context. Similarly, the schwa rule calculates an overall probability of the optional schwa being produced, regardless of its phonetic context either within or across words.

	Word Error
Original Dictionary	36.6%
Add all source pronunciations	33.9
Re-estimate probabilities	29.6
Phonetic Rules	30.1
Prune non-rule-based	29.4
Prune rule-based	29.8

Table 1: Results on BREF

Table 1 shows results for the various pronunciation modeling techniques described. The first result shows the system's original dictionary, which contains pronunciations from BDLex for the words in BDLex, and out-of-context TTS-derived pronunciations for all other words. The second result shows the benefit of introducing the in-context TTS-derived pronunciations. The third result uses a lexicon derived by re-estimating the probabilities for each pronunciation from a forced alignment (only one iteration of this process was performed; further iterations were found to have a negligible effect). In the fourth result, rule probabilities for liaison and schwa deletion are calculated, and rule-derived probabilities are merged with the empirically observed probabilities for each word. Finally, in the fifth and sixth results, pruning is applied to both the non-rule-derived and rule-derived re-estimated dictionaries. Here, all results shown use the same neural net, trained from an alignment generated from the re-estimated lexicon. This allows us to compare the dictionary used as the only variable across experiments. One might expect that this would put the original dictionary at a disadvantage, as the net is not tuned to that system. However, our experiments indicate that such a net in fact performs worse than the net trained using the re-estimated lexicon (38.2% w.e.r.).

As can be seen, the pronunciation modeling process gives a 6.8% absolute cumulative performance gain. Simply introducing a number of more reliable pronunciations gives us most of the improvement. Adding probability estimation helps further.

Using rule probabilities did not result in a significant change in performance. However, we believe that using rule probabilities is important because without them, if a given word happens not to be seen in the training data in, say, the form with a liaison consonant, that pronunciation will be eliminated entirely from the dictionary. Using rules ensures that all forms of the canonical dictionary pronunciation are kept with non-zero probability. We found an overall 13% chance of a liaison consonant being produced, and a 50% chance of an optional schwa being produced.

One reason that pruning helps is that the alignment process used to segment utterance pronunciations into word pronunciations is imperfect. Pruning eliminates most of the bad pronunciations introduced by this process.

4. DISCUSSION

One difficulty is designing speech recognition systems for French is the level of homophony. For example, the singular and plural of most nouns and adjectives are pronounced identically (except when liaison occurs) but spelled differently. While such ambiguities should in theory be resolved by the language model, present-day systems, including trigger or topic-based language models, are incapable of handling the syntactic phenomena involved. The fact that such cases are counted as errors is one of the main reasons that performance for the current system on BREF is much lower than on comparable English tasks.

Another peculiarity of French is the related phenomena of liaison (e.g. a [z] sound being pronounced in "les amis"), and elision (e.g. "le" becomes "l' " before a vowel. These are both cross-word effects — however our recognition system does not consider the context in which a word is found. (Another approach is described in [7].) We simply have two pronunciations for "les" in the dictionary [le] and [lez]. Similarly, "l' " and "le" are considered two different words by our system. Not being able to use context to rule out either the pronunciation with or without liaison has the potential to increase confusibility in the system. However, simply having both pronunciations in the dictionary works fairly well. It is interesting that our pronunciation modeling techniques originally designed for English, which do not take these factors into consideration, nonetheless gave a significant performance improvement.

5. THE CHOICE OF PHONEMES

Several interesting questions are raised in choosing a phoneme set for speech recognition for French. For example, we merged the [ɛ̃] and [œ̃] phonemes, based on reports in the phonetics literature that many speaker that not make the distinction, and the relatively small number of word pairs differentiated by the distinction. Another interesting question is how to handle the distinction between [e] and [ɛ] and between [o] and [ɔ]. We chose to investigate the [e]/[ɛ] distinction because of its importance in differentiating the very common words "et" and "est", as well as many other words such as imperfect, future, and conditional verb forms.

The pronunciations found in BDLex only distinguish [e] and [ɛ] in word-final position, using the symbol [ɛ] to transcribe both sounds in all other positions. Furthermore, "et" and "est" are homophonous, both transcribed as [ɛ]. This is in accordance with phonetic studies indicates that speakers do not systematically distinguish the two sounds. However, it goes against traditional rules of French pronunciation, as found in a standard dictionary, or a classroom. In the results reported above, we trained our system on pronunciations that always used the [ɛ] phoneme except word finally, despite the fact that, in BDLex, the symbol [ɛ] in non-final position actually means "[ɛ] or [e], or something in between".

In order test the claim that the two phones are not in fact distinguished (even in the relatively careful read speech of the BREF corpus), and in an attempt to improve the performance of our recognizer on confusible pairs such as "est" and "et", we trained the system using a new dictionary created by retranscribed all non-final instances of [ɛ] as either [e] or [ɛ] depending on the word's orthography. Sounds spelled as é were transcribed [e]; all other instances remained as [ɛ]. This is fairly accurate as a rule of "traditional" French pronunciation, although it does not capture such phenomena as [ɛ] raising to [e] before [i] [u] and [y], or é being pronounced [ɛ] in closed syllables such as found in "médecin". We also transcribed [e] as the only pronunciation for "et", and [ɛ] and [ɛt] as the only pronunciations for "est". After realigning the data with the dictionary thus produced, we retrained the neural net, and tested performance on the test data.

As shown in table 2, this did not in fact improve results. (These results are not comparable to Table 1 due to a different language model.) We interpret this as an empirical indication the [e]/[ɛ] distinction is not systematically made by French speakers, even in read speech.

	Word Error
Final Distinction Only	29.9%
Orthographic-Based Distinction	30.3
Merge into one Phone	32.3

Table 2: Comparison of Phoneme Sets

Given that a more refined distinction between [e] and [ɛ] does not improve results, we decided to try merging the two phones into one by retraining the net on data with only one output for both phonemes. This removes the inconsistency of using the [ɛ] phoneme for different acoustic meanings at the end of a word and within a word. As shown in table 2, this phone merger actually worsened our results. While the difference in error between lines 1 and 2 is not statistically significant, the increase from 30.3% to 32.3% is. Thus we conclude that maintaining two phonemes for [ɛ] and [e] is useful, despite the fact that they are only distinguished word-finally, and that the common words "est" and "et" are not distinguished at all.

6. CONCLUSION

Often in speech recognition, we find that the simple improvements give the greatest gain in performance. Thus our pronunciation modeling system was able to significant improve a French-language speech recognizer, despite its lack of sophistication as to the peculiarities of the language. The benefits of the system are fairly language-independent. We hope to continue this work by introducing cross-word pronunciation modeling and improving the language model to better handle the level of homophony found in French. Our recognition results support the counter-intuitive claim that [e] and [ɛ] are not distinguished within words by French speakers.

ACKNOWLEDGEMENTS: This work was performed at the Institut Dalle Molle d'Intelligence Artificelle Perceptive in Martigny, Switzerland, with the support of Hervé Bourlard. We would like to thank Nelson Morgan, Johan Andersen and Olivier Deroo for their assistance. This study was conducted with support from a NDSEG fellowship, NSF SGER grant IRI-9713346, the European Community basic research grant SPRACH, under a subcontract from the Faculté Polytechnique de Mons in Belgium, and the European Community basic research grant THISL, under a subcontract from Sheffield University in the United Kingdom.

7. REFERENCES

[1] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[2] O. Deroo, C. Ris, F. Malfrere, and T. Dutoit. Modèles hybrides et reconnaissance de la parole continue indépendante du locuteur en français. In *XXIIèmes Journées d'Etudes sur la Parole*, pages 401–404, Martigny, Switzerland, 1998.

[3] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. V. D. Vreken. The mbrola project: Toward a set of high quality speech synthesizers free of use for non commercial purposes. In *ICSLP-96*, page I., Philadelphia, Pennsylvania, 1996.

[4] M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Eurospeech-97*, 1997.

[5] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. RASTA-PLP speech analysis technique. In *IEEE ICASSP-92*, pages I.121–124, San Francisco, CA, 1992.

[6] L. Lamel and G. Adda. On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In *ICSLP-96*, Philadelphia, Pennsylvania, October 1996.

[7] L. Pousse and G. Perennou. Dealing with pronunciation variants at the language model level for the continuous automatic speech recognition of french. In *Eurospeech-97*, Rhodes, Greece, September 1997.

[8] S. Renals and M. Hochberg. Decoder technology for connectionist large vocabulary speech recognition. Technical report, Sheffield University Department of Computer Science, September 1995.

[9] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 109–116, Kerkrade, Netherlands, April 1998.

[10] T. Sloboda and A. Waibel. Dictionary learning for spontaneous speech recognition. In *ICSLP-96*, Philadelphia, Pennsylvania, October 1996.

[11] G. Tajchman, E. Fosler, and D. Jurafsky. Building multiple pronunciation models for novel words using exploratory computational phonology. In *Eurospeech-95*, Madrid, Spain, September 1995.

[12] P. Woodland, M. Gales, D. Pye, and S. Young. Broadcast news transcription using htk. In *IEEE ICASSP-97*, Munich, Germany, April 1997.