

A study of implicit and explicit modeling of coarticulation and pronunciation variation

Stéphane Dupont, Christophe Ris, Laurent Couvreur & Jean-Marc Boite

Multitel & FPMS-TCTS, Avenue Copernic 1, B-7000 Mons, Belgium

dupont,r,couvreur,boite@multitel.be

Abstract

In this paper, we focus on the modeling of coarticulation and pronunciation variation in Automatic Speech Recognition systems (ASR). Most ASR systems explicitly describe these production phenomena through context-dependent phoneme models and multiple pronunciation lexicons.

Here, we explore the potential benefit of using feature spaces covering longer time segments in terms of implicit modeling of coarticulation and pronunciation variants.

The study is based on the analysis at the phonetic level of the performance of context-independent and context-dependent acoustic models, and more particularly the impact of modeling different time context going from 70 ms up to 310 ms on typical cases of pronunciation variants.

Results, confirmed by word recognition experiment, put into light some ability of generic acoustic models to implicitly handle pronunciation variation.

1. Introduction

Current state-of-the-art speech recognition technology builds on Hidden Markov Models whose states are defined as context-dependent phonetic units, typically triphones. Triphones allow the estimated probability density functions of phonemes to depend on the nature of the previous and following phonemes [1]. Such models acknowledge the impact of coarticulation resulting from the mechanics of speech sounds production. Recent studies have also naturally attempted to permit this left and right context to span beyond the adjacent phonemes [2].

Several techniques are also attempting to explicitly model pronunciation variations that exist in natural speech. These are relying on expert knowledge in order to define phonologically pertinent pronunciation variations [3]. Other approaches are based on data-driven methods that build such pronunciation rules from statistical phonetics or data [4], or on a mix of data-driven and knowledge based methods [5]. Significant improvement are obtained in extreme cases of variations, like in non-native speech, for which few material is available for training purposes [6]. However, when the training data is representative of the task, only limited improvements in accuracy have been obtained to date, despite the sound underlying knowledge and methodologies. Part of the explanation of these modest results has been given by the fact that context-dependent phonetic units provide modeling of such variants in an implicit fashion [7].

Besides, other studies tackle the assumptions underlying the HMM stochastic framework. Some techniques have reached global adoption in current ASR technology. They are generally based on joint modeling of speech frames together with their temporal derivatives, or modeling of a sequence of speech frames. These techniques allow to capture the fine-grained

structures and temporal dependencies in the speech signal, and hence complement the coarse modeling of coarticulation provided by the triphone models. It has been recently suggested that such techniques could also provide some form of implicit modeling of coarticulation, that would make the use of triphones partly redundant, just like pronunciation variants might be partly superfluous in the case the system is equipped with context-dependent models.

In this paper, we contribute to this debate with some experiments where assumptions are gradually relaxed. From a context-independent (CI) modeling on single frames, we switch to context-dependent (CD) modeling, and to CI and CD modeling of larger time segments (section 2). As a first step, we decide to avoid focusing on a particular recognition task, which could reduce the diagnostic nature of our study. We rather work on phonetic recognition and study the ability of CI and CD acoustic models to implicitly model a set of identified pronunciation variations (section 3). For this purpose, we are using data from the French Bref corpus, that have been manually annotated at the phonetic level.

Finally, in section 4 speech recognition results are also presented on a speaker independent recognition task of 35,000 isolated words (very high perplexity).

2. Explicit and implicit modeling

In fluent speech, the acoustic realization of the phonemes is strongly influenced by the phonetic context. This phenomenon, called coarticulation, is due to the intrinsic inertia of the human speech articulatory system. In other words, the transition from one phoneme to another is achieved smoothly, in a continuous way, probably constantly optimizing the tradeoff between the vocal effort and the speech intelligibility. The influence of the phonetic context on the realization of the phonemes can even lead to pronunciation variants where the target phoneme can be withdrawn or a different phoneme can be pronounced. Moreover, regional accents are also inducing variations in pronunciation. These phenomena are often handled in ASR by introducing triphone models and providing multiple phonetic transcriptions of the vocabulary words. In the following sections, we describe the feature extraction and modeling techniques used in our experiments.

2.1. Feature extraction, acoustic and lexical modeling

The reference front-end is based on Perceptual Linear Predictive (PLP) processing [8] but MFCC processing would yield similar recognition performance.

On the acoustic modeling side, our speech recognition systems are based on the hybrid HMM/ANN architecture [9] where ANNs - typically multilayer perceptrons (MLP) - are used to

estimate the HMM state likelihoods. This collaboration between ANNs and HMMs has proven its efficiency on many different speech recognition tasks. In its "classical" form used in this paper, a single ANN is trained in order to classify the frames of acoustic features into language dependent acoustic units (phonemes, diphones, ...). In that framework, it can be shown that the outputs of the MLP approximate the local posterior probabilities of the acoustic units [9]. ANN architecture offers a direct way of introducing time context information by putting several feature vectors at its input. In our experiments, time segments going from 70 ms up to 310 ms will be modeled this way.

The hybrid system is one way to approach probability density function modeling, which may need less data than full generative modeling (like GMMs - Gaussian Mixture Models), and hence may show benefit of implicit models with reasonable amounts of training data. All our acoustic models have been trained on 100 hours of French read speech from the Bref corpus [10].

2.1.1. CI and CD acoustic modeling

Context-independent (CI) acoustic modeling refers to the classical hybrid HMM/MLP configuration where each output of the ANN corresponds to one phoneme. Including silence, we come with 35 outputs.

In the case of context-dependent (CD) acoustic modeling, the acoustic units are typically the diphones (phonemes in particular left OR right phonetic contexts) or triphones (phonemes in particular left AND right phonetic contexts). The extension of the hybrid HMM/MLP system to CD phonemes is not easy. Some approaches can be found in the literature but either lead to a crude approximation of CD models (combination of CI models and transitions in [11]) or imply very complex systems (combination of multiple ANNs in [12]).

In this paper, we have defined triphones as a concatenation of left and right diphones, coming from $34^3 = 39304$ triphones, to $2 * 34^2 = 2312$ diphones. Moreover, phonetic contexts have been clustered into 11 classes (based on phonological rules), reducing the number of CD units to $2 * (34 * 11) = 748$. Hence, the CD acoustic model consists in one MLP trained on those targets.

Note that both the CI and CD models have been designed for having approximately the same amount of parameters (about 1 million) and that these parameters are estimated on the same training corpus.

2.1.2. Explicit modeling of pronunciation variants

Many pronunciation variants can be predicted from the sequences of canonical phonemes by applying a set of coarticulation rules. Each phoneme can be depicted by a set of articulatory features, as voicing, degree of aperture, place of articulation, ... [13]. The inertia of the human speech production system suggests that these features change smoothly in time, possibly inducing pronunciation variants. The coarticulation rules integrate these articulatory constraints and predict potential alternative word/sentence pronunciations.

This set of rules can either be automatically extracted from acoustic data or described by expert knowledge of the language [5], as in our case.

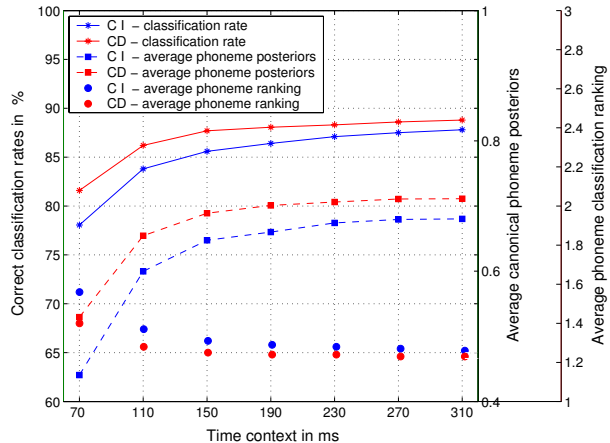


Figure 1: Phoneme modeling performance of CI and CD acoustic models according to the contextual information. Reference is the canonical transcription.

3. Phonetic analysis framework

In this section, we study the impact of the acoustic modeling on a phoneme classification task, going from CI, short time context (70 ms), to CD, long time context (310 ms) systems. As we expect more complex systems to implicitly model digressions from canonical pronunciations, we have used the canonical transcriptions as a reference. In order to have a finer analysis of the performance of different acoustic models, we not only evaluate phoneme recognition rates but also the average phoneme posterior probabilities. At the local level (frame or phoneme, compared to word or sentence), these measures give a more reliable idea of the intrinsic performance of the acoustic models than classification rates.

3.1. Phoneme modeling performance

Figure 1 shows the evolution of the performance of the CI and CD acoustic models according to the contextual information used to train those models (time windows spanning from 70 ms to 310 ms). The performance is evaluated by three measures (actually highly correlated):

1. the correct phoneme classification rate according to the canonical reference (in %)
2. the average posterior probability of the canonical phonemes regardless of classification errors. The posteriors are normalized so as to sum to one at each time frame.
3. the average ranking of the canonical phoneme in the N-best list.

The classification and ranking measures show that CI and CD models tend to converge in terms of performance when increasing the contextual information, which could demonstrate that CI acoustic models can make use of this larger context to implicitly model the coarticulation. However, this convergence is not as clear for the average posterior probability, which may suggest a still better potential for CD models in word recognition, as misclassified phonemes may appear with better posteriors.

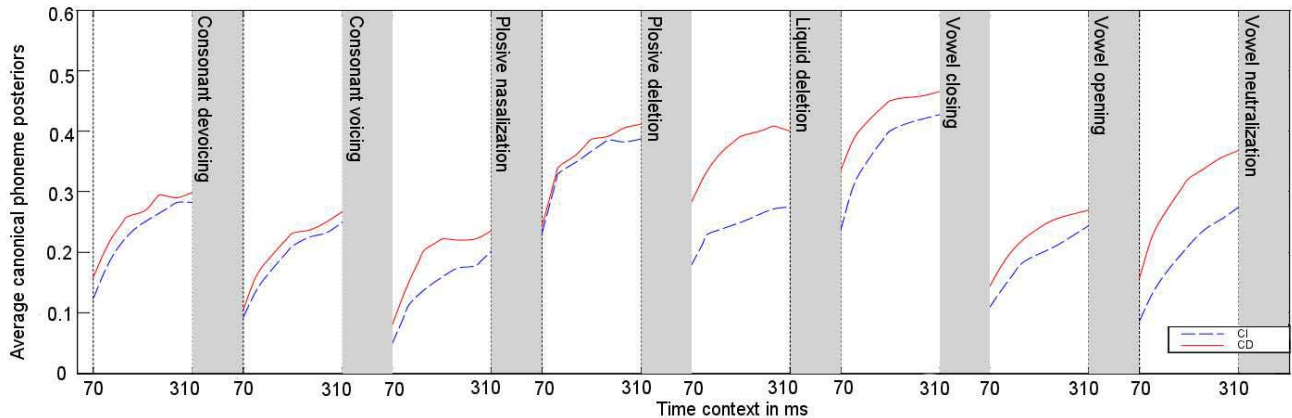


Figure 2: Performance evaluation of CD and CI acoustic models according to time context on 8 classes of pronunciation variations.

3.2. Analysis of pronunciation variants

In order to closely study the impact of pronunciation variants on the speech recognition system, we have manually annotated at the phoneme level 75 minutes of French read speech, that is about 40,000 phonemes. This so-called "surface" annotation corresponds to the sequences of phonemes that are considered to have actually been pronounced by the speakers. Differences between this "surface" transcriptions and the canonical transcriptions can be considered as pronunciation variants. We have defined 8 classes of such variants corresponding to the most frequent instances and covering about 5% of the annotated data and 60% of all the pronunciation differences in the surface transcriptions:

1. the devoicing of consonant is a particular case of sound assimilation. A voiced consonant is devoiced in the presence of an unvoiced sound. For instance,
 - *médecin* - / m E d s e ~ / → / m E t s e ~ /
 - *sauve qui* - / s o v k i / → / s o f k i /
2. the voicing of consonant is the opposite phenomenon.
 - *disgrâce* - / d i s g R a s / → / d i z g R a s /
 - *anecdote* - / a n E k d O t / → / a n E g d O t /
3. the nasalization of plosives is another case of assimilation. In the presence of nasal vowels, plosives may be nasalized.
 - *trente deux* - / t R a ~ t d 2 / → / t R a ~ n d 2 /
 - *demande pour* - / ... a ~ d p u ... / → / ... a ~ n p u ... /
4. The deletion of the plosives is a particular case of the reduction of the consonant groups.
 - *direct permet* - / ... E k t p E ... / → / ... E k p E ... /
 - *extrême* - / E k s t R E m / → / E s t R E m /
5. The deletion of the final liquids /l/, /R/ is due to the relaxation in final position, especially in presence of a plosive.
 - *capable* - / k a p a b l / → / k a p a b /
 - *reconnaître* - / R @ k O n E t R / → / R @ k O n E t /
6. The closing of vowel is a reduction of the aperture due to a vocalic harmonization, especially in case of open syllables.
 - *aimer* - / E m e / → / e m e /
 - *loger* - / l O z e / → / l o z e /
7. The opening vowel is the opposite phenomenon.
 - *sauf* - / s o f / → / s O f /
 - *taupe* - / t o p / → / t O p /

8. The vowel neutralization is the replacement of an open vowel by the neutral vowel /@/. This is a sort of relaxation too.

- *international* - / ... a s j O n a l / → / ... a s j @ n a l /
- *gouvernement* - / g u v E R n ... / → / g u v @ R n ... /

The modification of the degree of aperture (case 6 and 7) is also characteristic of regional and foreign accents.

Figure 2 shows the evolution of the average posterior probability of the canonical phonemes according to time context varying from 70 ms to 310 ms, for the eight identified pronunciation variants described above.

We first notice that the posterior values are significantly lower than the average values depicted in Figure 1, which demonstrate the real weakness of the acoustic modeling for those phenomena. Here again, we note the strong benefit of incorporating larger temporal windows and further demonstrate that some sort of implicit modeling of pronunciation variations is actually achieved. Indeed, we see that even deletions of phonemes may be partly handled by the acoustic model. The effect of the CD modeling is more questionable according to this Figure. We note, indeed, a big improvement for certain pronunciation variations such as "liquid deletion" or "vowel neutralization" while others like "consonant voicing/devoicing" or "plosive deletion" seem to not take advantage of the CD modeling. This could be due to the way CD models are built; in order to reduce the number of parameters, the different phonetic contexts are clustered. The way these clusters are settled may reduce their ability to handle particular pronunciation variants. These issues probably deserve further investigations.

We have to note that the modeling of several consecutive frames (up to 31 in our case) relaxes, in some way, one of the underlying HMM assumptions, namely the output independence assumption, which states that within a particular HMM state, the vectors that the model emits are statistically independent and identically distributed [1]. This naturally seems to be a strong assumption and it is still not well understood to what extent relaxing it would provide much gain, and what it would imply in terms of modeling structure complexity and training data needs.

4. Speech recognition evaluations

Speech recognition experiments have also been performed to confirm and illustrate the enhanced accuracy of both implicit and explicit contextual modeling. The results reported here

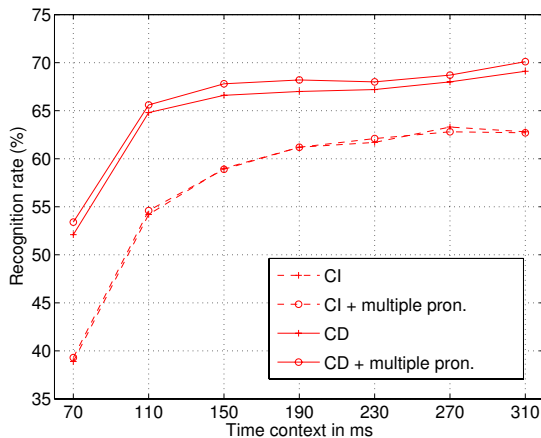


Figure 3: Recognition rates using CI and CD acoustic models on 35,000 words.

all concern ASR for the French language, and use the systems described previously, that have been trained on 100 hours of French read speech.

We draw on the baseline feature set used above, but additive and channel noises are also handled using a combination of Wiener filtering and temporal trajectory filtering.

The ASR experiment is related to an isolated word recognition task in French. The task concerns a dictionary query application where the user is allowed to pronounce any word within a dictionary of 35,000 entries. As no language model is used, the performance heavily depends on the acoustic modeling. The test corpus contains nearly two thousand samples of isolated words collected from 25 speakers. The background noise level is moderate.

Figure 3 presents the word recognition rates obtained using four systems: a) CI acoustic models + baseforms dictionary, b) CD acoustic models + baseforms dictionary, c) CI acoustic models + multiple pronunciation dictionary, d) CD acoustic models + multiple pronunciation dictionary.

The baseform phonetic transcriptions are those of the Brulex dictionary [14]. The multiple pronunciations have been obtained by processing the baseforms using contextual rewriting rules defined by a linguist (see section 2.1.2). The average number of variants per word is 6.1. Results show that improvements discussed in the previous section (and obtained on data matched to the training corpus i.e. same task, same recording conditions) translate to a totally different task.

5. Conclusions

In this paper, we had a close look on the actual ability of the ASR acoustic models to handle the coarticulation and pronunciation variation phenomena in fluent speech. Starting from standard techniques of explicit modeling of these phenomena, namely, the triphone modeling and the multiple pronunciation dictionaries, we gradually switch to fully implicit modeling through modeling of feature spaces covering longer time segments. Phoneme modeling experiments demonstrate the potential of acoustic models to implicitly handle different types of pronunciation variants. Experiments on word recognition show a good consistency of the improvement across tasks and recording conditions, hence, demonstrating the improved genericity of the proposed modeling.

Of course, the results we present in this paper are very dependent on the approaches used for building the CD models and the multiple pronunciation dictionaries and should be validated on more complex tasks including spontaneous speech. The task perplexity should be studied as an important parameter as conclusions on the performance at the word level may differ depending on the task itself, as well as the language perplexity.

Acknowledgements

This work has been partly supported by the EU 6th Framework Programme, under contract number IST-2002-002034 (project DIVINES). The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein. We are grateful to Pascale Woodruff for putting together the large vocabulary isolated words database and task and to Sophie Roekaert for designing the multiple pronunciation rules for French.

6. References

- [1] X. Huang, A. Acero & H. Hon, "Spoken Language Processing: A Guide to Theory, Algorithm and System Development", Prentice Hall PTR 2001.
- [2] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev and S.J. Young, "The 1994 HTK large vocabulary speech recognition system" in Proc. of ICASSP'95, Detroit, 1995.
- [3] T. Hazen, L. Hetherington, H. Shu & K. Livescu, "Pronunciation Modeling Using a Finite-State Transducer Representation", in Proc. of PMLA'2002, Colorado, Sep. 2002.
- [4] M. Magimai-Doss & H. Bourlard, "On the Adequacy of Baseform Pronunciations and Pronunciation Variants", IDIAP-RR 04-27, Switzerland, 2004.
- [5] H. Strik & C. Cucchiarino, "Modeling pronunciation variation for ASR: a survey of the literature", Speech Communication, 29:225-246, 1999.
- [6] K. Bartkova & D. Jouvet, "Multiple Models for Improved Speech Recognition for Non-Native Speakers", Proc. of SPECOM'2004, St. Petersburg, Russia, Sep. 2004.
- [7] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiyang & Z. Sen, "What Kind of Pronunciation Variation is Hard for Triphones to Model?", in Proc. of ICASSP'01, Salt Lake City, Utah, 2001.
- [8] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", The Journal of the Acoustical Society of America, vol.87, nr.4, april 1990, pp. 1738-1752
- [9] H. Bourlard and N. Morgan, "Connectionist Speech Recognition: A Hybrid Approach", Kluwer, 1994.
- [10] L.F. Lamel, J.-L. Gauvain and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French", in Proc. of Eurospeech'91, pp. 505-508, 1991.
- [11] S. Dupont, J.M. Boite, C. Ris, O. Deroo, V. Fontaine & L. Zaroni, "Context Independent and Context Dependent Hybrid HMM/ANN Systems for Training Independent Tasks", in Proc. of Eurospeech'97, Rhodes, Greece, pp. 1947-1950, 1997.
- [12] H. Franco, M. Cohen, N. Morgan, D. Rumelhart & V. Abrash, "Context-dependent Connectionist Probability Estimation in a Hybrid Hidden Markov Model / Neural Net Speech Recognition System", in Computer Speech and Language, n°8, pp. 211-222, 1994.
- [13] M. Debrock and P. Mertens, "Phonétique générale et française. Une introduction.", Louvain : Presses Universitaires, 1990.
- [14] F. Content, P. Mousty and M. Radeau, "BRULEX: Une Base de Données Lexicales Informatisée pour le Français Ecrit et Parlé", in L'Année Psychologique, pp. 551-566, 1990.