SPEECH RECOGNITION BY GOATS, WOLVES, SHEEP and … NON-NATIVES

Dirk Van Compernolle
Lernout & Hauspie Speech Products NV
Koning Albert I Laan 64, 1780 Wemmel, Belgium
Tel. +32 2 456 05 00, Fax +32 2 460 01 72, E-mail Dirk.VanCompernolle@lhs.be

ABSTRACT

This paper gives an overview of current understanding of acoustic-phonetic issues arising when trying to recognize speech from non-native speakers. Regional accents can be modeled by systematic shifts in pronunciation. These can often better be represented by multiple models, than by pronunciation variants in the dictionary. The problem of non-native speech is much more difficult because it is influenced both by native and spoken language, making a multi-model approach inappropriate. It is also characterized by a much higher speaker variability due to different levels of proficiency. A few language-pair specific rules describing the prototyical nativised pronunciation was found to be useful both in general speech recognition as in dedicated applications. However, due to the nature of the errors and the mappings, non-native speech recognition will remain inherently much harder. Moreover, the trend in speech recognition towards more detailed modeling is counterproductive for the recognition of non-natives.

INTRODUCTION

That recognition of non-native speech is significantly harder than that of native speech can't be a surprise. We as humans often have a hard time understanding someone speaking his second or third language. We might also readily determine the accent and will quickly make an assessment on the degree of non-nativeness.
But we also know that there is not something like "a non-native". French, Japanese and Indians will speak English in a very different way. The sounds will not just be accented, but they will insert and delete phonemes, they will make grammatically weird sentences, etc. After some time we may get used to the peculiarities of their speech and understand them quite well. Listening to another non-native in a language, non-native for ourselves, sometimes turns out not to be too difficult because the speaker uses a restricted vocabulary and easy syntax.

A speech recognizer is often compared to a person who is bad of hearing, a young child or to someone who isn't too familiar with the language. So maybe a recognizer should like non-native speech. We'll see that this is not at all the case as the recognizer will take little or no advantage from the reduced grammatical complexity, but will suffer greatly under miserable acoustic phonetic conditions. So a recognizer will only see the bad sides of non-native speech and generally poor robustness of speech recognition systems will show double.
In this review paper I will focus on the acoustic phonetic issues. It is structured as follows. First I'll discuss native accents; then I will revisit the complexity of differences in phoneme spaces across languages, moving on to the complexity of non-native speech recognition for general purposes and dedicated applications.

ACCENTS AND DIALECTS

CHARACTERIZING ACCENTS

Each living language has numerous accents which are continuously on the move. It's sometimes implicitly assumed accents will differ most distinctively in the realization of vowels[Bary89], but consonantal differences may be strong as well. Eg regional distinctions in Latin American Spanish are especially pronounced for a few consonants.
Accents will only show minor differences at the higher - abstract - phonemic level, but the specific acoustic-phonetic realizations might shift considerably. Small phonetic shifts can freely be applied to almost all sounds of any language without having any impact on recognition as all languages only use a limited part of the articulatory space. As phonemic ambiguity shouldn't increase markedly by accent shifts, a strong shift of one class could have a forceable impact on other classes as well. It is possible that accents introduce or remove homonym confusions, but overall acoustic confusability should not change significantly.
In terms of pattern recognition one might describe an accent as a shift in classes across the feature space, but with maintenance of the same degree of separability of the classes. Typical of native accents is that these shifts

will be applied in a pretty consistent manner by whole groups of speakers.

There have been two main paths in attacking the dialect problem for speech recognition. The first one tries to to model accents as pronunciation variants at the detailed phonetic level[Bary89,Cohe89,Adda98]; the other one doesn't get involved with detailed modeling but creates multiple models for large speaker groups [VCom91,Beat95,Drax97].

Existence of accents questions the validity and feasibility of symbolic representation of sounds, but at the same time highlights the tremendous abstraction applied in our alphabetic writing systems. At the abstract (phonological) level a unique symbolic representation may suffice for a whole group of accents. If, on the contrary, we want to represent all the different realizations in a symbolic (phonetic) way, then the better chance is that no system will be detailed enough. Straightforward reasoning also leads to a few more conclusions. Because of the continuity of the shifts that are feasible at the pronunciation level, any symbolic representation is inherently local and not universal. Abstraction and symbolic representation are hence not absolute but relative and only valid within the applicable language. Phoneme boundaries aren't absolute, but defined wrt. to the collection of phonemes valid for that language. Ultimately it follows directly from the continuity of the characteristic sound shifts, that granularity and categorization of dialects is a very ill-defined problem.

Now, let's confront the above hypotheses with experiences with real world speech recognition. The Dutch/Flemish language group is an interesting case study as accent and dialect diversity is tremendous, given its compact geography, but we'll restrict to the case where everyone at least attempts to speak the "standard" language and not the local dialect. Contrary to the British/US English distinction there are no spelling differences between Dutch and Flemish.

MODELING ACCENTS BY MULTIPLE ACOUSTIC MODELS

Everyone who has tackled the problem of Dutch/Flemish speech recognition knows that models trained on one group perform very poor on the other group. Error rates may double or triple. Relaxing within class variability will not help, because it isn't random extra variability that needs to be modeled, but a systematic shift. Putting all data in a single model gives reasonably satisfying results, but will still be significantly (eg20%) worse than accent specific models. There are also some interesting asymmetries showing increased or decreased separability for certain classes depending on the accent. One such example are the digits. For Dutch speakers the pair 'twee/τωε˥/-drie/δρι/' (similar as for German zwei-drei), while for the Flemish the pair 'vijf /ϖΕιφ/ -zes/ζεσ/' is by far the more confusable one. The above can be understood by following two characteristic differences of Dutch vs Flemish:

- Diphtongization of long vowels by Dutch, reduces the ee-ie phonetic distances. This goes together with a stronger diphtongization of the real diphtongs in Dutch vs Flemish which increases the distance of ei-e
- Devoicing of voiced fricatives, which is stronger however for the /v/ than for the /z/ which increases the phonetic distances of the v-z pair.

Interesting to note is that the above shifts get more pronounced the further north one goes and that the geographical boundary for these phenomena might even be better characterized by the Maas-Rijn Delta than the Belgian-Dutch border.

The strength of the shifts - up to the phonemic level - causes a strong overlap of distributions in a global modal, while accent specific distributions are much better separated. The latter may be a good criterion to decide if accents should be modeled as extra speaker variability in a single model or if multiple models are required. The above is also a good example that accent shifts can either somewhat reduce or enhance phonetic contrasts between words. These small changes may have little impact on human performance, but show up in machine based recognition.

Now that usage of 2 models for Dutch/Flemish seems perfectly reasonable, one may wonder how many more models would make sense and how to define them. In some early work on this problem [VCom91] it was found that extra models based on regional clustering provided little or no advantage, but the interpretation may have been influenced by insufficient data to train a larger number of models. In unrelated more recent work, it was found that 3-4 models does make sense.

In similar experiments for US English [Beat95], it was found that 3 accent models for the US gave a good tradeoff between performance, compactness and trainability of the models.

Overall we can conclude that using multiple models for the different dialects is an easy and effective way to improve performance. Modeling of a very small number of well designed large clusters seems to perform better than many small clusters, because of loss in intrinsic speaker variability in the clusters when insufficient training data is available.

MODELING ACCENTS BY PRONUNCIATION VARIANTS

The strong phonetic differences between Flemish and Dutch or British and US English would intuitively suggest another way to model strong accent differences, i.e. by pronunciation variants[Cohe89]. In last year's

ESCA workshop on pronunciation variation much interesting work was presented [eg Adda98,Rile98], but often with somewhat disappointing results. Only the most pronounced variants are essential, especially so for the most frequent short words of a language. When modeling variants in great detail, eg for speaking style differences, then increased confusability seems to offset the increased modeling capacity.
A major weakness of implementing accent variability by multiple pronunciations in a single dictionary is that accent consistency for a given speaker is not enforced. Therefore, another approach - which is rarely feasible in real-time speaker independent systems - is the use of parallel phonetic dictionaries, with dictionary selection on a maximum likelihood criterion. This is easily done however in speaker dependent and/or speaker adaptive dictation systems where the choice can be based explicit speaker preference or after parallel batch processing.

In the speech recognition world British and US English are most often treated as 2 different languages with different spellings, separate phonetic dictionaries - probably even different phonetic alphabets. It comes somewhat more intuitive than in the Flemish/Dutch case because of the spelling differences and the geographical separation. Nevertheless, it can be shown that speech recognition performance will still be very reasonable if the phonetic baseforms from one variant are used for the other, but trained with the correct speaker group. It shows great resilience of phonetic transcriptions against accent variation as long as the canonical transcription only needs to be valid for a coherent regional group and not for multiple groups at the same time. This is explained by the fact that most pronunciation variants will be learned implicitly when building context dependent acoustic models.

CROSS-LINGUAL PHONETICS

IPA AND ITS COMPUTER EQUIVALENTS

Alphabetic writing systems must stand out as one of the greatest inventions of all times. It made it possible to write about every language with as few as 30 symbols, corresponding to the sounds of the language. Due to independent evolution of the Roman alphabet in different languages and further emphasized by the independent evolution of written and spoken language, the phonetic consistency is far away in most of today's languages and complicated grapheme-2-phoneme converters are necessary to go from written to spoken language.
Modern phonetic alphabets are in a way a reinvention of the original alphabet and try to write according to the rule "one sound - one symbol ". The IPA (International Phonetic Alphabet) is the concerted international effort that tries to achieve this (illusive) goal for all languages of the world at once. That each language only sparsely fills the articulatory and acoustic space is well illustrated by the fact that the IPA needs several hundred basic symbols to encompass all languages. Several ASCII compatible computer derivatives are used by speech community has derived its own derivatives (SAMPA, Worldbet). At L&H we developed our own version L&H+ for internal usage. These cross-lingual phonetic alphabets greatly enhance readability but at the same time create the false impression (hope) of the existence of a truly language independent phonetic alphabet.
Extensive experience over the past 5 years in speech technology applications has shown how illusive the target "one sound - one symbol" might be. L&H+ foresees in about 300 different classes for the 30 odd languages that it is currently used for. Despite all efforts and good definitions, there remains a great lack of inconsistency between transcriptions in different languages. This is due to the enforcement of a single symbol on multiple classes which are close but not truly identical. One of the complicating aspects is that no phonetician exists who can claim native or close to native pronunciation for a sufficiently large group of languages. Thus even the best implementation is based on a consensus of experts who don't really understand each other.

LISTENING AND SPEAKING BY NON-NATIVES

There are many similarities but also a few significant differences between accents of natives and pronunciations of non-natives. Class definitions are only valid within a single language (and accent) and there is no reason whatsoever why class definitions of one set should be portable to another one. The very fine distinctions will get lost in any compact symbolic representation. Similarly some of those distinctions we do hear and others we don't. Which distinctions we hear, depends much on our language exposure at younger age. It's not so extreme that we have learned strict class boundaries applicable only to our native language, but it seems that we have learned to listen for sound features which are most relevant to our native language[Fox95], somehow projecting all acoustic features onto a lower dimensional space appropriate for our native tongue. And by feedback mechanisms our acoustic and articulatory spaces are tightly coupled, so we only pronounce those sounds adequately that we need in our native tongue.
Numerous straightforward examples can be given. The tonal phonetic features of oriental languages are tough to hear and learn for Europeans because it didn't get engraved in their front end acoustic processor. Somewhat less pronounced, but well demonstrated, is that natives of different European languages might discriminate vowels along different feature dimensions [Fox95]. Thus what is a phonemic distinguishing feature for a native of one

language may hardly be audible to a native of another one. Consequently you must expect that a non-native will significantly mispronounce sounds that are not in his native auditory collection, by projecting the pronunciation onto his own articulatory and acoustic space. As an example, don't be surprised if you hear a Spanish person mention *'a shit of paper'*, by omission of the duration cue in the word *sheet*. Similarly, I shouldn't be too surprised if both human and machine recognizers mistakes my 'p' for a 'b' by lack of aspiration of the 'p'. While the aspiration is a distinguishing feature in English it is not in Flemish, where it does not exist.

Thus there are significant differences between native and non-native accents. Native accents are all based on pretty much the same phoneme set. Because of proximity, it is reasonable to assume that acoustic feature space and distinguishing acoustic clues will be very similar and the average phonemic contrast will be maintained across native accents. Native accents are information preserving transformations. Non-natives will project sounds onto a subspace defined by the intersection of target language and native language, thus on an inherently lower dimensional feature space, thus potentially with loss of information. And the further that languages are apart from each other, the worse the intersection will be and the greater the information loss [Bona98].

MULTI-LINGUAL SPEECH RECOGNITION

Our inherent skepticism about cross lingual phonetic alphabets can be put to test by a multi-lingual speech recognition system.

In recent years, several groups have tried to build cross language phone models. The ultimate goal would be that one sufficiently large collection of phoneme models is sufficient to model all the languages of the world. But more often the goals are more restrictive. It is either used to have a compact footprint for multilingual systems or to bootstrap or augment the training of acoustic models in a new language when little data is available [Köhl96,Bona97,Schul98].

At L&H we've also used such systems to deal with initial responses in a multi-lingual system with a priori unknown language by the caller. This avoids the problem of ranking scores between 2 systems with completely different models. The results we found are similar to the ones found elsewhere in the literature.

- Multilingual phone models perform worse than single language phone models, provided there is enough training data for each of the languages
- The effect becomes more pronounced as more diverse languages are grouped together. This is naturally explained on the basis that phoneme classes from far away languages cluster intrinsically less good, but it may also be a hint that the multi-lingual phonetic alphabet misses some important details.
- Degradation may be on the order of 20-80% depending on the number and diversity of languages that are clustered.
- Despite their poorer performance, such systems may have a high practical value, especially when little or no data exists in a particular language or in some simple but intrinsic multilingual tasks

NON NATIVE SPEECH RECOGNITION

MORE DATA OR DIFFERENT MODELS ?

Based on the above, the easy way out might be to consider non-native speech as just another (heavy) accent. If the occasional pronunciation errors are modeled as random then we can even forget about them. So all we need is data. To some extent it is a valid approach, except that … variability is much larger and non-natives are by no means a homogeneous group. At least the influence of the native language needs to be taken into account. Thus, if we need to start collecting data on non-natives, then the whole data collection problem becomes quadratic in nature and is clearly not feasible nor can it be the right approach. Here we are just running into the limits of more and more data. Assuming that the data problem is quadratic might even be underestimating the real dimensionality. It is well known that people talking in their third, fourth .. language might copy - correctly and incorrectly - pronunciations from other foreign languages they know. All of this is further complicated by the large variability in language proficiency among the non-natives.

So is there anything else to do than lay back and observe that non-natives are worse than natives ? Digging deeper, the situation looks even more grim. Much of the progress in the last 15 years in acoustic modeling is based on more detailed modeling, by creating sharper and sharper distributions for narrower and narrower classes. This is diametrically opposite of the tolerance and robustness required for non-natives. Distribution of non-native scores on allophonic variants will greatly differ from the distribution of natives, because they will emphasize different cues. So it should come as no surprise that for people with heavy accents the performance gain between context-independent and context-dependent models might get totally washed out.

SPEAKER ADAPTATION FOR NON-NATIVES

There is another feature about non-natives which has significant impact on ASR systems. Vocabulary of non-natives tends to be much more limited and occurrence of

unknown words will not be uncommon. These are likely to happen in enrollment scripts. Whenever an unknown word occurs, the speaker will hesitate and apply certain letter-2-sound rules, typically a mix of the rules of his native tongue mixed with the non-native one, leading to all kind of funny pronunciations.

Potential for speaker adaptation will thus greatly depend on proficiency of the non-native. If all words in the adaptation script are known to the non-native, then we fall back to the 'thick accent' case. If there are many unknown words, hesitations will occur and gross mismatches between pronunciation and transcription will be present. Such mismatches will not shift the sound categories to their desired location, but will randomly smear out the distributions. One way to avoid this is to include only speech with minimal confidence levels, but as could be expected, this is even more difficult for non-natives. For reasonably proficient non-natives, speaker adaptation has shown dramatic improvements[Zava95] reducing the error rate by a factor 2-3 without adaptation of the phonetic baseforms. This confirms the assumption that a very strong accent shift needs and can be modeled by transformation of the distributions. However, even after adaptation, non-natives still performed a factor 2 worse than natives. This is explained by a combination of effects: (i) random pronunciation errors and (ii) projection of pronunciation onto a lower dimensional, less discriminative, space. Another more subtle cause may be that the chosen state tying - necessary in speaker adaptation - is optimized for natives and might be less applicable to the non-native accents.

NATIVISED PRONUNCIATIONS

Pronunciation errors are common with unknown words, and even more so if simple letter-2-sound rules are insufficient as is the case for proper names - a common problem in Europe with its density of languages and high mobility. The two most immediate application areas are automated attendants and car navigation.

The automated attendant in our office is a good example of how complex a small problem quickly gets. There are roughly 100 employees of whom about 60% are Flemish natives of whom most but not all have a name with Flemish pronunciation. The only other significant language group are the French speakers. In total there are names of 12 different language origins of which 4 from outside Europe. Despite the monolingual English greeting, the name pronunciation might be in many different ways, given in order of occurence: native pronunciation, pronunciation with a Flemish accent, pronunciation with an English accent, pronunciation with another accent. This is in stark contrast to the implementation of similar systems in US or France, where almost all users would have a tendency to bastardize the name pronunciations to the local language. Given the great mix of pronunciation and accent, there is no option for a language-pair specific solution and one needs to rely on some "language independent" recognizer as the symbol set from a single language will be insufficient to code all the various transcriptions that one might require. On average 2-3 transcriptions of each name suffice to yield acceptable performance. Given the sparseness of the language mix, we did not make great attempts to derive general rules that would describe prototypical pronunciation variants. The system has been operational internal for several years and many of us have learned fail safe pronunciations for the names we often use.

Another case is the one of car navigation, as explored in the EC VODIS project[VODIS]. Assume a German travelling to France and talking to the navigation unit in German while specifying French location names.

It was found that Germans - also the ones with little French knowledge - have some knowledge of French phonology and ultimately use a mix of French and German letter-2-sound rules[Tran99]. A reasonable approximation of the real pronunciations is obtained by starting from the correct French pronunciation and applying a small set of French-2-German conversion rules. Most of these can be related to the absence of a very close relative of a particular sound in the native language.

While done in an ad hoc manual way, part of the above work can be automated and common mutations could be learned on the basis of a moderate body of German pronunciation of French names. At the same time it becomes obvious that many similar rules - but maybe somewhat reduced - would apply for a German speaking French. Similar rule based work has been reported in the field of nonnative speech recognition [Bona98], pronunciation variation in general [Crem98]. Today, this may stand out as one of the more promising approaches in dealing with non-native pronunciations.

LANGUAGE LEARNING

One of the most extensively researched topics in non-native speech recognition is the one of language learning[Stil98]. For this application there may be many more novice speakers than others who have already a thorough knowledge. The most intuitive measure to evaluate someone's pronunciation is some form of confidence measure. But similarly as with native speech recognition, simple likelihood measures aren’t a most reliable metric, and it's correlation with expert ratings was found to be low[Neum96]. It was found that rate-of-speech [Cucc98,Neum96] is a reliable estimator of degree of non-nativeness. However, ROS has little diagnostic value as it does not identify pronunciation errors.

Likelihood scores can be turned into a much more reliable measure if they are turned into a likelihood ratio of speaker vs. prototypical native pronunciation. In order to obtain a reference score pronunciations of 10-20 native speakers of all sentences in a lesson can be recorded and processed by the recognizer. This procedure has been found to yield significantly better performance than the use of more generic methods to generate the reference score in the likelihood ratio.

Still an alternative approach for turning likelihood ratios into indicators of pronunciation errors, is the explicit modeling of expected errors. Due to the very different phonotactic structure of Japanese vs. English, many pronunciation errors made by Japanese, learning English, can be predicted[Kawa98]. Consonant clusters, which are non existent, will lead to vowel insertions and diphtongs are likely to be replaced by a single vowel. A pronunciation network including the correct and incorrect pronunciations is subsequently fed to the recognizer and simple Viterbi alignment shows immediately all errors. The latter approach is very efficient for the small group of frequent language-pair specific errors. Basically the same set of rules applies as discussed in the previous section on nativised pronunciation.

CONCLUSIONS

In this paper we reviewed the difficulties arising when recognizing non-native speech, especially the additional difficulties compared to dealing with native accents. Non-natives are more complex than heavy accented speakers. Across different applications it was found that a few language pair specific rules can describe many of the typical mispronunciations. However, because the loss of certain distinguishing acoustic cues and heavy shifts in pronunciation, non-native recognition will be very difficult for today's recognizers using sharp distributions. It stresses the inherent lack of robustness of our current acoustic-phonetic modeling. Likelihood scores should gracefully decay as phonetic feature distance grows which is not necessarily the case in a state of the art recognizer.

ACKNOWLEDGEMENTS

Much of the discussion presented in this paper is based on unpublished work at L&H from Jose Conejo, Kristin Daneels, Bart D'Hoore, Luc Mortier, Louis ten Bosch, and Filiep Vanpoucke.

REFERENCES

[Adda98] M. Adda-Decker, L. Lamel, "Pronunciation variants across systems, languages and speaking styles", Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, pp.1-6, Rolduc, May 1998.

[Bary89] W.J. Bary, C.E. Hoequist and F.J. Nolan, "An approach to the problem of regional accent in automatic speech recognition", Computer Speech and Language, 3, pp.355-366, 1989.

[Beat95] V. Beattie et. al. "An integrated multi-dialect speech recognition system with optional speaker adaptation", Proc. Eurospeech95, pp.1123-1126.

[Bona97] P. Bonaventura, F. Gallocchio, G. Micca, “Multilingual speech recognition for flexible vocabularies”. Proc. Eurospeech’97, pp 355-358, 1997

[Bona98] P. Bonaventura, F. Gallocchio, J. Mari, G. Micca, "Speech recognition methods for non-native pronunciation variants", Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, pp. 17-22, Rolduc, May 1998.

[Cohe89] M. Cohen "Phonological Structures for Speech Recognition", Ph.D. Thesis, UC Berkeley, 1989.

[Crem98] N. Cremelie, J.P. Martens "In search for pronunciation rules", Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, pp. 23-28, Rolduc, May 1998.

[Cucc98] C. Cucchiarini, F. de Wet, H. Strik and L. Boves "Assessment of Dutch pronunciation by means of automatic speech recognition technology", Proc. ICSLP 98, pp.751-754.

[Drax97] C. Draxler and S. Burger, "Identification of regional variants of high German from digit sequences in German telephone speech", Eurospeech 97, pp.747-750.

[Fox95] R.A. Fox, J.E. Flege and M.J. Munro, "The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis". JASA Vol.97(4), pp. 2540-2511, 1995.

[Kawa98] G. Kawai, K. Hirose "A method for measuring the intelligibility and Nonnativeness of phone quality in foreign language pronunciation training", Proc. ICSLP 98, pp.782-785.

[Köhl96] J. Köhler "Multilingual phoneme recognition exploiting acousitc-phonetic similarities of sounds" Proc. ICSLP96, pp.2195-2198.

[Neum96] L. Neumeyer, H. Franco, M. Weintraub and P. Price "Automatic text-independent pronunciation scoring of foreign language student speech" Proc. ICSLP 96, Philapdelphia 1996, pp.1457-1460.

[Rile98] M. Riley et. Al. "Stochastic Pronunciation modelling from hand-labelled phonetic corpora", pp109-116, Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, May 1998.

[Schul98] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Large Vocabulary Speech Recognition", Proc. ICSLP98.

[STIL98] ESCA ETRW Workshop **STiLL** Speech Technology in Language Learning ,May 25-27 1998, Marholmen, Sweden

[Tranc99] I. Trancoso, C. Viano, I. Mascarenhas and C. Teixeira "On deriving rules for nativised pronunciation in navigation queries", EUROSPEECH 99

[VCom91]D. Van Compernolle, J. Smolders, P. Jaspers and T. Hellemans "Speaker Clustering for Dialectic Robustness in Speaker Indpendent Recognition", Eurospeech91, pp.723-726

[VODIS] VODIS, “Advanced Speech Technologies for Voice Operated Driver Information Systems”, EC Language Engineering Project LE 1-2277.

[Zava96] G. Zavagliakos, "Maximum A Posteriori Adaptation For Large Scale HMM Recognizers", Proc. ICASSP96, pp. 725-728