LEARNING PRONUNCIATION AND FORMULATION VARIANTS IN CONTINUOUS SPEECH APPLICATIONS

*D. Colibro, L. Fissore, C. Popovici, C. Vair ^, P. Laface**

^ Loquendo, Torino, Italy
(Daniele.Colibro,Luciano.Fissore,Cosmin.Popovici,Claudio.Vair}@loquendo.com

*Politecnico di Torino, Italy
Pietro.Laface@polito.it

ABSTRACT

Most voice driven applications are based on recognition grammars. In complex applications it is difficult to exactly predict how the users will formulate their requests even if a careful study of the user's behavior has been performed. Moreover, it is possible that a speaker's word pronunciation does not match the phonetic transcription of the system, mainly in the case of foreign words.
Loquendo has developed a tool that collects field data, detects the most significant weaknesses of the application due to pronunciation of formulation mismatches, and filters the collected field corpora. This permits the application designers to perform their analysis only on a reasonable amount of pre-processed and automatically labeled data.
This paper presents the approaches that have been devised to detect pronunciation variants of vocabulary words and linguistic formulations not covered by the recognition grammar. Results showing the improvements that have been obtained including automatically detected formulations in three grammars for two languages are also detailed.

1. INTRODUCTION

In complex applications it is difficult to exactly predict how the users will formulate their requests even if a careful study of the user's behavior has been performed.
The a-priori knowledge provided to the system is useful to issue a first release of a speech application, but this is not enough for its success. The system should be able to adapt its grammar to the formulations of the users accessing the service, and to the phonetic transcriptions of the vocabulary words [1]. A related issue is how to deal with the pronunciations of words from non-native or strongly accented speakers [2].
The main source of information that can be used for these tasks is field data. However, the efforts required to label at the sub-word level huge amounts of collected data by hand, and to perform a screening of them would make this impractical for human operators. Active learning approaches, see [3] for example, could be used to elegantly solve these problems.
Loquendo has developed a tool - currently available for all the languages covered by Loquendo ASR - that collects field data, detects the most significant weaknesses of the application due to pronunciation or formulation mismatches, and filters them. This tool, thus, allows the application designers to perform their analysis only on a reasonable amount of pre-processed, and automatically labeled data. The application log collects all the information related to a single recognition interaction: the recognized words, their confidence values, the constrained and unconstrained phonetic transcriptions, and, if required, the audio signal.
This paper presents the approaches that have been devised to detect vocabulary words pronunciation variants and linguistic formulations not covered by the recognition grammar. This work extends to grammars the phonetic learning approach that has been successfully applied, for isolated word recognition, to an automated Directory Assistance system [4]. This service, developed for Telecom Italia, is operational from the year 2000. It deals with both business and residential requests from a database of 25 million Italian subscribers.
The main observations that have motivated this work are:

- Poor confidence scores can be used as word or grammar mismatch indicators
- Different utterances having the same content produce similar phonetic transcriptions
- Partitioning the field data into phonetically similar clusters allows detecting user formulations or pronunciations not covered by the application.

The paper is organized as follows: Section 2 gives a short overview of Loquendo system. Section 3 illustrates the confidence measures used to detect signal regions of acoustic/phonetic mismatch. The generation of lists of candidate pronunciation and formulation variants is detailed in Section 4. Section 5 presents the strategy for selecting the variants that are used to update the system knowledge. Experimental results and our conclusions are given in Section 6 and 7 respectively.

2. PHONETIC DECODING

Loquendo ASR is a recognizer based on a Hybrid HMM-NN model, where the emission probabilities of the HMM states are estimated by a Multi Layer Perceptron. It is able to use both language models and grammars. The decoder uses a set of units modeling the stationary parts of the context independent phonemes (less affected by the phonetic context), and a larger set of transition units defining all the transitions between the stationary units that can be reliably trained.

Table 1– Results of the phone-looped model recognizer

Lang	# pho	Phone Accuracy	Del rate	Ins rate	Sub rate
it-it	27	80.0%	5.1%	5.2%	9.7%
es-es	32	76.9%	5.9%	4.4%	12.8%
en-us	45	62.7%	7.1%	9.7%	20.5%
en-gb	47	54.4%	6.4%	10.5%	28.4%
de-de	48	52.9%	5.5%	13.9%	27.7%

The system also produces, together with the grammatical constrained word hypotheses, the *free phonetic transcription*, i.e. the best sequence of phones obtained using a phone-looped model.
The accuracy of the phonetic decoder has been evaluated on the same training corpora that have been employed for estimating the acoustic models for the languages available with the Loquendo ASR. The phone accuracy, the deletion, insertion and substitution error rates, obtained aligning the free phonetic transcriptions with their references, are shown in Table 1, for a subset of these languages. It is worth noting that the phone accuracy is inversely related to the number of phones defined for each language.

3. CONFIDENCE SCORING

To select useful information from field data, without knowing the corresponding word transcriptions, we need a measure of the reliability of the recognition results. Our approach does not rely on application level information, such as user confirmations or human operator support. Even if the application information is valuable, we avoid its use. For this reason the tool is not bound to a given application design and can be used in any context.
The reliability measure that we use to detect regions of acoustic mismatch is an acoustic confidence score, *ALLR(w,)* based on posterior probability estimates of local phones, generated by the hybrid HMM/NN model [5]

$$ALLR(w) = \frac{\sum_{n=b}^{e} \log P(s_w^* \mid O_n)}{\sum_{n=b}^{e} \max_{s \in S} \log P(s_w \mid O_n)} \qquad (1)$$

where w is a word, b and e are its beginning and ending frames according to the Viterbi segmentation, S is the set of output states of the NN model, O_n is the *n-th* acoustic observation vector, and s^*_w is the sequence of states produced by the Viterbi alignment for word w.
ALLR(w) is, thus, the ratio between the free score, given by the sum of the a posteriori log probability of the best matching state for each frame, and the sum of the frame scores constrained by the model of word w. This measure is easily obtained in a hybrid HMM/NN model because all the posterior probabilities are computed in parallel by the NN. The values of *ALLR(w)* range from 0 to 1, and the maximum is reached when the free score and the constrained one for each frame are the same, indicating an optimal acoustic matching according to the model. Low values of *ALLR(w)* are, instead, good indicators of acoustic mismatch. The confidence measure of (1) can also be used to compute the accuracy of a hypothesized phone, rather than the accuracy of a word.

4. PHONETIC LEARNING

When collecting a large number of utterances referring to the same grammatical context, clusters of phonetically similar strings can be obtained. The central elements of the most significant clusters are quite accurate phonetic transcriptions of (possibly new) user formulations or pronunciations.
The similarity between two phonetic transcriptions is evaluated by Viterbi alignment of the two strings using, as local distance, the log-probability of insertion, deletion and confusion among phones. These probabilities are trained by aligning each canonical phonetic transcription of the training database with its corresponding free phone transcription.

4.1. Pronunciation variants

The detection of possible pronunciation variants for a grammar's word requires the collection of a set of utterances related to that word. Since a word can be embedded in a sentence, the free phonetic transcription corresponding to the temporal boundaries of a decoded word is considered as an instance of pronunciation of that word. The free phonetic transcriptions are collected in different sets (one set per word) and used in the clustering process described in section 4.3.
Since the decoded word sequence can be inaccurate, we insert in a word set only the free phonetic transcriptions, related to the instances of that word, recognized with a medium-high confidence. This assumption is reasonable for medium size grammars where the acoustic confidence is a good measure of correct recognition, because the acoustic confusability among grammar words is typically quite low. For large grammars, a human check is necessary to validate the consistence of a word set. The check aims at avoiding that the phonetic transcriptions included in the set are related to different words.
Finally, it is worth noting that task of learning pronunciation variants requires a good quality of the baseform transcription of the grammar words. Loquendo ASR relies on the high quality phonetic transcriber that is also used by the Loquendo TTS synthesizer.

4.2. Formulation variants

In the Directory Assistance application described in [3], the task of learning formulation variants has been addressed, for isolated words, by collecting and clustering the phonetic strings corresponding to user requests that the automatic system was not able to complete.
The same approach has been extended in this work, to generic grammar directed tasks. In this scenario, our goal is to detect the utterance regions that are not well covered by the recognition grammars. A naïve approach would detect low confidence sequences of words. The corresponding free phone transcription sequence can then be added to a set, labeled "unknown", and all the sequences in the "unknown" can be clustered to detect new formulation variants.
This approach has the drawback that it is unable to detect a new formulation whose temporal boundaries are not equal to the decoded words boundaries. Our solution to this problem is to compute a frame level (instantaneous) confidence measure for the phone sequence corresponding to the decoded words. The frame confidence is computed as a running window moving average of the phones confidence. The windows length is of the order of 50 frames (each frame lasts 10ms).
Table 2 shows the free and grammar constrained phonetic transcription of the utterance "The state of Indiana, thank you" (REF), recognized by a simple grammar, *covering only isolated US states and city names*. The decoded word (REC) is "Indianapolis".
Figure 1 shows the instantaneous confidence for free and constrained phonetic transcriptions of this utterance.
Using the confidence score of the detected words, we obtain the three segments <sil>, "Indianapolis", and <sil> as shown in Table 2.

Table 2– Free and grammar constrained transcriptions of the sentence “The state of Indiana, thank you”

REF	Frames	Free	Constrained	REC
<sil>	50	<sil>	<sil>	<sil> conf=0.02
T H E	62	Dh		
	67	i		
S T A T E	74	s		
	85	t		
	90	HEI		
	107	p		
O F	115	HEh		
	120	f		
I N D I A N A	134	`i:	i	I N D I A N A P O L I S conf=0.43
	141	n	n	
	152		d	
	154	Hj		
	158		i	
	165	HEh	HEh	
	180	n	n	
	187	Ae	Ae	
T H A N K	202	f	p	
	220	Ae	HEh	
	224		l	
	226		i	
	230	N		
	235		s	
	238	k		
Y O U	247	Hj	<sil>	<sil> conf=0.11
	256	`u:		
<sil>	263	<sil>		

The first and the third segment have low confidence scores, less than 0.4, but while the first one correctly covers the utterance “The state of”, the last one does not match “Thank you”. On the other hand, using the instantaneous confidence we correctly detect both the first and last formulations that are not covered by the grammar.
A list of potential new formulations, not foreseen by the grammar designer, can be obtained clustering these data.

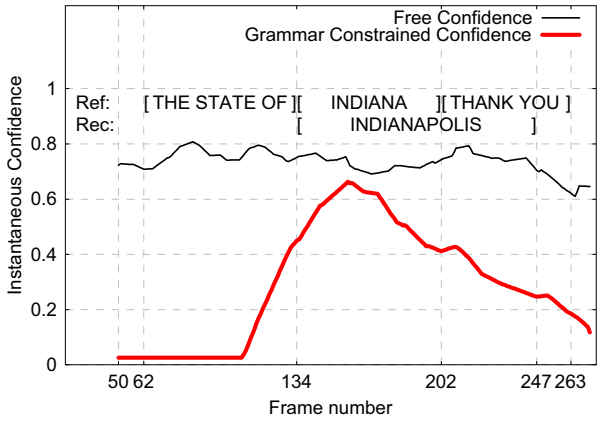
4.3. Clustering

The free phone transcriptions collected in the previous steps and associated to a grammar word set are clustered into similar subsets to produce a list of possible new *pronunciation* variants. Clustering the transcriptions in the “unknown” set generates, instead, a list of candidate *formulation* variants.
The subsets are created using a furthest neighbor hierarchical cluster algorithm, based on the mutual distance between each phonetic string. To compute the distance among a huge number of phonetic strings, we use a recursive tree to tree matching procedure, where a tree branch is a phonetic transcription [4]. Since we are interested in clusters with a small dispersion of the included elements, we ignore transcription distances greater than a small threshold. This dramatically reduces the cost of the matching procedure.

5. SELECTION OF NEW VARIANTS

Significant clusters are characterized by a high cardinality and small dispersion of the included phonetic strings. The central element of a significant cluster is the one that achieves the minimum value for the sum of distances to all the other elements. Clusters with few elements and large within cluster variance are discarded. The central elements of clusters with high cardinality and small dispersion represent phonetic transcriptions that can be inspected as possible new formulations or pronunciation variants. This information can be exploited to update the corresponding grammar or word transcriptions. In particular, a rule that has been found useful for adding new pronunciation variants is that a new transcription is added if the cardinality of its cluster is greater than the 10% of the collected phonetic transcriptions for that set. Before being included in the system, a hypothesized transcription is compared – using Viterbi alignment – with the transcriptions already in the system, to avoid increasing the confusability among similar sounding words.
Adding new transcriptions or formulations is a responsibility of the application developers. They can check the produced hypotheses accepting or refusing them on the basis of their semantic knowledge of the application. When audio recording has been enabled, a candidate transcription can be checked against one or more samples of the pronunciations that have been clustered to produce the proposed variant, listening to only a small amount of selected data.

Figure 1– Instantaneous confidence for the free and grammar constrained phonetic transcriptions of the same utterance



6. EXPERIMENTAL RESULTS

A number of tests have been carried out to assess the effectiveness of the described phonetic learning technique.
For pronunciation variants learning, the experiments have been performed on three built-in grammars – “Currency”, “Date”, and “Time” – for the UK English and German languages. The goal of these tests was to verify the performance improvement after the insertion of the detected pronunciation variants. Table 3 shows the number of words and the number of added variants for the three tested grammars.

Table 3– Size of test grammars and number of added variants

	English UK		German	
	words #	added variants #	words #	added variants #
Currency	118	54	122	68
Date	158	89	242	90
Time	111	31	126	27

Table 4– Word Accuracy for English UK language

Grammar	Training			Test		
	# utt	baselineWA	learn.WA	#utt	baselineWA	learn.WA
Currency	3652	82.8	83.1	916	75.9	78.7
Date	7260	84.9	85.6	1630	77.1	78.5
Time	7016	90.2	91.3	1737	87.9	88.9

Table 5– Word Accuracy for German language

Grammar	Training			Test		
	# utt	baselineWA	learn.WA	#utt	baselineWA	learn.WA
Currency	3347	94.0	94.3	1031	91.9	92.1
Date	7277	90.8	91.8	2106	89.8	90.4
Time	6439	94.7	94.8	1818	91.9	92.1

The speech corpus that has been used to learn the pronunciation variants is a subset of SpeechDat 2, while part of the SpeechDat Mobile corpus has been used for testing. The training data has been used to learn new formulations and to perform preliminary recognition tests. The test data, collected in a mobile phone environment – rather than in the PSTN environment of the training – has been used to validate the learning approach.
Table 4 and Table 5 compare the baseline word accuracy and the word accuracy obtained after the insertion of the new pronunciation variants derived for the three grammars both in UK English and in German, evaluated on the Training and Test data sets. The average error rate reduction on test data set is 8.8% for English UK and 3.9% for German.
The small relative error rate reduction is significant for at least two reasons:

- Phonetic learning allows to improve the recognition performance even for grammars including common usage words
- The phonetic learning approach does not require any acoustic model retraining. The insertion of the new formulations has no additional maintenance cost but their insertion into the grammar

The assessment of the formulation variants learning has been done using an artificial test scenario because it is important to use labeled speech corpora and appropriate grammars to evaluate the quality and the number of the unforeseen formulations found. The goal of one of these learning tests is to detect the pronunciations of the days of the week within date expressions,
But these utterances are recognized by a grammar that does not cover the days of the week. The learned transcriptions, generated by processing 2336 date expressions in UK English, are shown in Table 6, where <# ele> is the set cardinality.

Table 6– New formulation variants

New Formulation	# ele	Guessed word (baseform)
m Ah N d HEI	312	Monday (m Ah n d HEI)
f Hr HAI d HEI	184	Friday (f Hr HAI d HEI)
T$ Hj `u: z d HEI	179	Tuesday (t Hj `u: z d HEI)
s Ah N d HEI	178	Sunday (s Ah n d HEI)
Hw e n z d HEI	139	Wednesday (Hw e n z d HEI)
s Ae t HEh d HEI	137	Saturday (s Ae t HEh d HEI)
t Hw e N t i	113	twenty (t Hw e n t i)
Th HAU z n d	107	thousand (Th HAU z n d)
t `u: Th HAU z n	95	two thousand (t `u: Th HAU z n d)
Th OR: z d HEI	88	Thursday I(Th OR: z d HEI)

All the seven days of the week have been correctly detected with very good transcriptions. A common error is the substitution of nasal \n\ by velar \N\ in *Monday*, *Sunday* and *twenty*. The detection of the in-grammar word *twenty*, *thousand*, and of the sequence of words *two thousand* is caused by the errors induced by the grammar that does not cover the day of the week formulations. As an example, "*Wednesday*, January *twenty,* nineteen fifteen" is recognized as "*Twenty-eight*, January nineteen fifteen". This happens because the day frame has been set to *28*, substituting word *Wednesday* not foreseen by the grammar, and *twenty* is not permitted anymore by the grammar constraints, because the day frame is already set.

7. CONCLUSIONS[1]

We have presented the techniques implemented by a tool recently developed by Loquendo to support the acquisition of field data useful for the diagnosis of the most significant application weaknesses, related to pronunciation or formulation mismatches. The tool enables application designers or maintainers to perform their analysis on a reasonable amount of pre-processed and labeled data only. The tool provides lists of candidate pronunciation or formulation variants, different enough with respect to the words or sequence of words covered by the application grammar. A candidate phonetic transcription can be inspected and easily related to the constituent words, possibly listening to a few samples of the pronunciations that have been clustered to produce the proposed variant.
The tool is particularly useful for detecting, from the field data, the actual pronunciation of words - for example city names - that are often inaccurately generated by a phonetic transcriber.
The obtained results, including automatically detected pronunciations in three built-in grammars, are promising if we consider that the canonical transcriptions of the words in those grammars are accurate. Moreover, the improvement holds also on channel mismatch conditions, without acoustic model retraining.

8. REFERENCES

[1] G. Williams, "Knowing What You Don't Know: Roles for Confidence Measures in Automatic Speech Recognition," PhD Thesis, Sheffield University, 1999.

[2] R. Goronzy, S. Rapp, R. and Kompe , "Generating non-native pronunciation variants for lexicon adaptation", *Speech Communication*, 42, (2004), pp 109-123.

[3] D. Hakkani-Tür, G. Riccardi and A. Gorin, "Active Learning For Automatic Speech Recognition", *Proc. of ICASSP-2002*, Orlando, pp. 3904-3907, May. 2002.

[4] C. Popovici, M. Andorno, P. Laface, L. Fissore, M. Nigra, and C. Vair, "Learning New User Formulations in Automatic Directory Assistance", *Proc. of ICASSP-2002*, Orlando, pp. 1-10, May. 2002.

[5] M. Andorno, P. Laface, and R. Gemello, "Experiments in Confidence Scoring for Word and Sentence Verification," *Proc. ICSLP-2002*, Denver, pp. 1377-1380, Sept. 2002.

[1] This work was partially supported by the EU FP-6 IST Project DIVINES – Diagnostic and Intrinsic Variabilities in Natural Speech