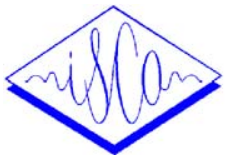

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

AUTOMATIC DERIVATION OF MULTIPLE VARIANTS OF PHONETIC TRANSCRIPTIONS FROM ACOUSTIC SIGNALS

Houda Mokbel and Denis Jouvet
France Télécom - CNET/DIH/RCP
2 avenue Pierre Marzin, 22307 Lannion cedex, France
e-mail: {mokbelh, jouvet}@lannion.cnet.fr

ABSTRACT

This paper deals with two methods for automatically finding multiple phonetic transcriptions of words, given sample utterances of the words and an inventory of context-dependent subword units. The two approaches investigated are based on an analysis of the N-best phonetic decoding of the available utterances. In the set of transcriptions resulting from the N-best decoding of all the utterances, the first method selects the K most frequent variants (Frequency Criterion) , while the second method selects the K most likely ones (Maximum Likelihood Criterion).

Experiments carried out on speaker-independent recognition showed that the performance obtained with the "Maximum Likelihood Criterion" is not much different from that obtained with manual transcriptions.

In the case of speaker-dependent speech recognition, the estimate of the 3 most likely transcription variants of each word, yields promising results.

1 INTRODUCTION

New words or Out-of-Vocabulary words are a major source of recognition errors for a speech recognition system. In real-world applications of speech tecnologies, it becomes essential to process unknown words. It involves detecting them, which then improves the performance of the system, and, if necessary, adding them to the system lexicon (in order to expand it). The twofold problem of OOV words has already been investigated and reported in literature. Detecting these words is in itself a difficult task. The basic technique for detecting new words consists in using a generic new-word model along with the models for the vocabulary words [1], [2] and [3]. This model should be detected whenever a new word occurs.

While the ability to automatically detect new words is important, it is also desirable to add them to the system lexicon. This way they can be recognized when encountered again. For adding a new word to the lexicon of a phonetically based speech recognition system, its phonetic transcription is necessary. This transcription can be obtained from the orthographic spelling by using a dictionary or a text-to-speech system [4], or from the orthographic spelling and a pronunciation [5]. The spelling of a new word is often unknown. However we can usually have one or several utterances of the word and an inventory of subword units as in [6] and [7]. In these two papers, a "single" phonetic transcription of the word was produced by maximising a likelihood. In [6], the subword units used were the fenones, while in [7], they were the phonemes.

In spontaneous speech, using a single pronunciation per word does not necessary yield the best recognition performance. It is therefore preferable to add alternative pronunciations to the phonetic dictionary, in a way that they model the given occurrences of words in the database. The approach proposed in [8] consisted in using sample utterances of words for generating phonetic transcriptions for them. The algorithm selected the "single" best transcription for each utterance of the word in the database, and computed a statistic of the resulting phonetic transcriptions of each word. Statistically relevant variants were added to the dictionary.

This paper describes two methods which use sample utterances of words for finding multiple transcription variants for them. The two methods investigated extend those of [8] and [7] respectively. The novelty lies in the selection of multiple variants from the N-best phonetic decoding for each utterance of the word. N transcriptions for each utterance were provided by the N-best algorithm, then k variants were selected from the resulting transcription set. Two selection criteria were developed and compared. These will be described in the following.

2 PHONETIC TRANSCRIPTIONS

2.1 N-best phonetic decoding

Let $y^{(1)}, ..., y^{(n)}$ be n given utterances of a word w, and let S be the set of all possible subword unit sequences. The N-best phonetic transcriptions $T_1^{(i)}, ..., T_N^{(i)}$ of the utterance $y^{(i)}$ are given by:

$$T_1^{(i)} = argmax_{s \in S} P(y^{(i)}/s)P(s) \quad (1)$$

$$T_2^{(i)} = argmax_{s \in S - \{T_1^{(i)}\}} P(y^{(i)}/s)P(s) \quad (2)$$

$$\vdots$$

$$T_N^{(i)} = argmax_{s \in S - \{T_1^{(i)}, ..., T_{N-1}^{(i)}\}} P(y^{(i)}/s)P(s) \quad (3)$$

The set of transcriptions of the utterance $y^{(i)}$ is:

$$\tau^{(i)} = \{T_j^{(i)}; j = 1...N\}$$

$P(s)$ is the $a\ priori$ probability of the subword unit sequence s.

Similarly, $\tau^{(i)}$ was determined for every utterance $y^{(i)}; i = 1, ..., n$. The $\tau^{(i)}$ are not disjoint; the same transcription can occur in the N-best transcriptions of several pronunciations $y^{(i)}$. We assume that the "good" transcription appears more frequently in the transcription sets. The first criterion of selection thus chooses the best transcription of a word according to the frequency of occurrence in the transcription sets of all the pronunciations of that word, (e.g. in $\{\cup\tau^{(i)}\}$). The second criterion is more rigourous, it is based on the maximum likelihood criterion.

2.2 Frequency Criterion

Let τ_w=$\{\cup\tau^{(i)}\} = \{T_j^{(i)}; j = 1, ..., N; i = 1, ..., n\}$ be the set of transcriptions of all the pronunciations of a word w. For this Frequency criterion, T is the "best" transcription of w if its frequency of occurrence in τ_w is maximal. Multiple phonetic transcription variants are obtained by keeping the K most frequent ones.

The following table shows the example of the digit "9" (in French "n.oe.f"), with its transcription variants found using the algorithm.

Table 1: **Transcription Variants for "9".**

τ_w, with w="9", $n = 5$, $N = 5$				
$\tau^{(1)}$	$\tau^{(2)}$	$\tau^{(3)}$	$\tau^{(4)}$	$\tau^{(5)}$
n oe f	n oe f e	n oe in f	**n oe f**	n o f
i n oe f	n oe f u	n in f	n a f	**n oe f**
ge oe f	n oe v f e	l in f	n e a f	n an f
i z oe f	n oe f b	**n oe f**	n oe in f	n a f
i l oe f	n oe v f u	n oe in s	n oe un f	n a in f

The transcription printed in bold face (**n oe f**) is the most frequent one. In this case, the most frequent transcription is the correct one (as it could be found in a lexicon). Unfortunately, this is not always the case.

2.3 Maximum Likelihood Criterion

In this method, the "best" transcription of w is the most likely one in τ_w, given the utterances of w. In other words, it is the one which is most likely to produce all n utterances, or the one for which $P(T/y^{(1)}, ..., y^{(n)})$ is maximal.

$$\hat{T} = argmax_{T\in\tau_w} P(T/y^{(1)}, ..., y^{(n)}) \quad (4)$$

$$= argmax_{T\in\tau_w} \frac{P(y^{(1)}, ..., y^{(n)}/T)P(T)}{P(y^{(1)}, ..., y^{(n)})} \quad (5)$$

$$= argmax_{T\in\tau_w} P(y^{(1)}, ..., y^{(n)}/T)P(T) \quad (6)$$

Assuming that the acoustic realisations are independent and that all the transcriptions are equivalent in the restricted set τ_w (e.g. the $a\ priori$ probability $P(T)$ is constant),

$$\hat{T} = argmax_{T\in\tau_w} P(y^{(1)}/T)...P(y^{(n)}/T) \quad (7)$$

$$= argmax_{T\in\tau_w} \prod_{k=1}^{n} P(y^{(k)}/T) \quad (8)$$

To obtain multiple transcription variants of w, we keep the K most likely ones.

The likelihoods are computed within the N-best decoder. If a transcription is not detected for a given utterance, the likelihood of the N^{th} transcription obtained for this utterance is associated to it.

2.4 Phonotactic constraints

Alternatives of the two algorithms consisted in imposing some phonotactic constraints on the loop of context-dependent phoneme models, in order to limit the possible successions of phonemes in the N-best phonetic decoding. This amounts to specify $a\ priori\ P(s)$ in the equations of §2.1.

Two kinds of constraints were adopted. The first one was a simple syllabic form allowing all sequences of the form:

$$Silence.(Syl) * .Silence$$

where $(Syl)*$ denotes a loop of syllables of the following structure:

$$\left(\begin{pmatrix}() \\ Cluster \\ Consonant \\ Semivowel\end{pmatrix}.Voyelle.\begin{pmatrix}() \\ Cluster \\ Consonant \\ Semivowel\end{pmatrix}.\begin{pmatrix}() \\ Pause\end{pmatrix}\right);$$

In the second constraint adopted, we used in parallel and between two "Silence", two other syllabic forms. Three different consonant clusters, the Start Cluster, the Middle Cluster and the Final Cluster (e.g. clusters occurring frequently at the start, in the middle and at the end of words respectively) were distinguished.

Note that we referred to the French syllabic construction. The consonant clusters used are the most frequent according to the French dictionary DELA and to the French lexicon BDPHO (oral speech) [9].

3 EXPERIMENTS

3.1 Speaker-independent recognition

The two approaches were first evaluated for speaker independent speech recognition, on several French databases recorded over the telephone network. The results reported correspond to a 36 French-word corpus. Experiments were carried out in order to measure the speech recognition performance as a function of the number of transcription variants considered in the lexicon, and of the number of sample utterances employed for determining these variants.

Experiments were conducted using PHIL90, the speech recognizer developed at CNET and based on Hidden Markov Models [10]. Figure 1 reports the error rates for both methods. For these experiments, we used the generic

looped model without any constraint on the phoneme sequences.

It turns out that without vocabulary-dependent training, the "Maximum Likelihood Criterion" outperforms the "Frequency Criterion". Using several variants (here 3) of pronunciation in the lexicon provides better performance than using a single one. In addition, in the case of 3 transcription variants, the performance obtained with the acoustically-based automatic transcription method using the "Maximum Likelihood Criterion", is comparable to that obtained with a manual transcription.

For these small vocabulary tasks, a vocabulary-dependent training compensates for possible incorrect transcriptions.

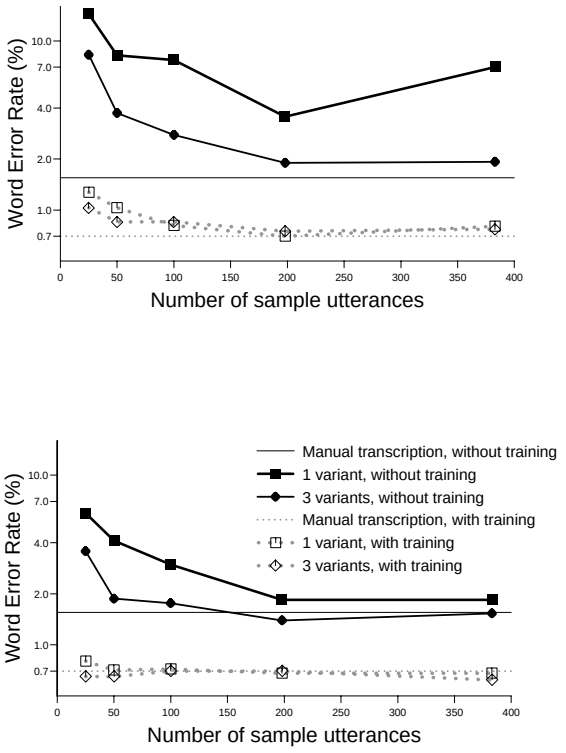

Figure 1: **Word Error Rate vs. number of sample utterances: Frequency Criterion (top) and Maximum Likelihood Criterion (bottom), without (solid lines) and after (dotted lines) vocabulary-dependent training.**

In other experiments, we examined the contribution of the phonotactic constraints to the recognition performance.. We found that the phonotactic constraints bring more with the "Frequency Criterion". Figure 2 shows the word error rates vs the number of transcription variants, before and after application of the first phonotactic constraint. The distinction between the "Clusters" (second constraint) does not significantly affect the results obtained with the first constraint.

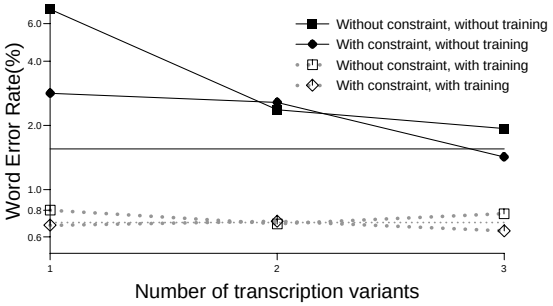

Figure 2: **Word Error Rate vs. number of transcription variants: Frequency Criterion, without (solid lines) and after (dotted lines) vocabulary-dependent training.**

3.2 Speaker-dependent recognition

The "Maximum Likelihood Criterion" was also evaluated for speaker-dependent speech recognition. The vocabulary used in this case was composed of 95 words pronounced by 40 speakers (20 males and 20 females). For each speaker, 3 or 4 training utterances per word were collected over the PSN telephone network. Speaker-dependent phonetic transcriptions were derived from these utterances and incorporated into the speaker's specific lexicon. Two other utterances per word were used for measuring the speech recognition performance.

To judge the performance of our automatic transcription method, the results were compared with those obtained using the fixed variance speaker-dependent word model. The use of the "Linear Multiple Regression" (LMR) adaptation procedure [11] was also investigated. The main results are reported in Figure 3. In this figure, the cumulated number of speakers is plotted as a function of the total recognition errors made. Point "A" means that 27 speakers have made 0 errors. Point "B" means that 33 speakers have made 0 or 1 error (< 2), etc. The higher the curve, the better the recognition performance.

The curves show that the simple estimate of the 3 most likely transcription variants of a word, without retraining of (speaker-independent) acoustic models, yields promising results. In addition, in the case of a speaker-dependent task, where the amount of training utterances is low (3 or 4), a speaker adaptation procedure based on "linear regression" is appropriate for adapting the model parameters to each speaker. This procedure led to a performance comparable to that obtained using the fixed variance speaker-

dependent word model.

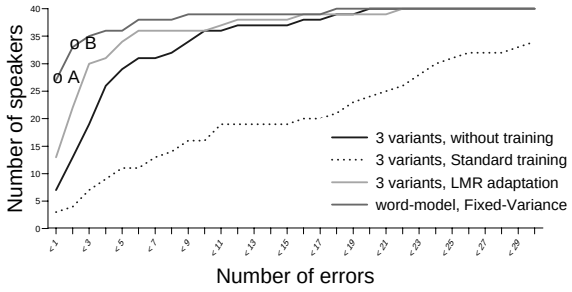

Figure 3: **Cumulated number of speakers vs. number of errors.**

4 CURRENT WORK

In our experiments, the number of transcription variants included in the lexicon was the same for all the vocabulary words (1, 2 or 3 variants). Algorithms are under investigation to determine the optimal number of transcription variants for each word.

5 CONCLUSION

This paper describes two methods for automatic transcription of words from acoustic signals. The method based on "Maximum Likelihood Criterion" gives a better performance than the Frequency based method. This performance is comparable to that obtained with a manual transcription in the case of speaker-independent recognition. The experiments showed that this method could also be applied to speaker-dependent tasks.

References

[1] A. Asadi, R. Schwartz and J. Makhoul. Automatic Detection of New Words in a Large Vocabulary Continous Speech Recognition System. In *ICASSP*, volume 1, pages 125–128, Albuquerque, USA, April 1990.

[2] S.R Young and W. Ward. Learning New Words from Spontaneous Speech. In *ICASSP*, volume 2, pages 590–591, Minneapolis, USA, April 1993.

[3] T. Kemp and A. Jusek. Modeling Unknown Words in Spontaneous Speech. In *ICASSP*, volume 1, pages 530–533, Atlanta, USA, May 1996.

[4] A. Asadi and H.C Leung. New-Word Addition and Adaptation in a Stochastic Explicit-Segment Speech Recognition System. In *ICASSP*, volume 5, pages 642–645, Minneapolis, USA, April 1993.

[5] A. Asadi, R. Schwartz and J. Makhoul. Automatic Modeling for Adding New Words to a Large-Vocabulary Continuous Speech Recognition System. In *ICASSP*, volume 1, pages 305–308, Toronto, Canada, May 1991.

[6] L.R Bahl, P.F Brown, P.V De Souza, R.L Mercer and M.A Picheny. A Method for the Construction of Acoustic Markov Models for Words. *IEEE Transactions on Speech and Audio Processing*, 1(4), pages:443–452, October 1993.

[7] R. Haeb-umbach, P. Beyerlein and E. Thelen. Automatic Transcription of Unknown Words in a Speech Recognition System. In *ICASSP*, volume 1, pages 840–843, Detroit, USA, May 1995.

[8] T. Sloboda. Dictionary Learning: Performance Through Consistency. In *ICASSP*, volume 1, pages 453–456, Detroit, USA, May 1995.

[9] V Auberge, L.J Boe and J.P Lefevre. Lexiques et Groupes Consonantiques. In *Journées d'Étude sur la Parole*, volume 1, pages 55–60, Nancy, France, September 1988.

[10] D. Jouvet, K. Bartkova and J. Monné. On the Modelisation of Allophones in an HMM Based Speech Recognition System. In *EUROSPEECH*, volume 2, pages 923–926, Genova, Italy, September 1991.

[11] T. Soulas, C. Mokbel, D. Jouvet and J. Monné. Adapting PSN Recognition Models to the GSM Environment by Using Spectral Transformation. In *ICASSP*, volume 2, pages 1003–1006, Munich, Germany, April 1997.