A NEW APPROACH TO SPEAKER ADAPTATION BY MODELLING THE PRONUNCIATION IN AUTOMATIC SPEECH RECOGNITION

Florian Schiel

Lehrstuhl für Datenverarbeitung,
Technische Universität München, Germany

ABSTRACT – To deal with large lexica (more than 2000) many systems of automatic speech recognition (ASR) use an internal phonetic representation of the speech signal and phonetic models of pronunciation from the lexicon to search for the spoken word chain or sentence. Therefore there is the possibility to model different pronunciations of a word in the lexicon. In German language we observed that individual speakers pronounce words in a typical way that depends on several factors as: sex, age, place of living, place of birth, etc. Our goal is to enhance speech recognition by automatically adapting the models of pronunciation in the lexicon to the unknown speaker. The obvious problem is: You can't wait until the present speaker will have uttered approx. 2000 different words at least one time. We solved this problem by generalization of observed rules of differing pronunciation to not observed words.

Another point presented is speaker adaptation by re-estimating the a-posteriori probabilities of the phonetic units used in a 'bottom up' ASR system. A word hypothesis is evaluated by the product of the a-posteriori probabilities of the phonetic units produced by the classification to the phonetic units belonging to the word hypothesis. Normally these probabilities are estimated during the training of the ASR system and stay fixed during the test. We propose a algorithm which observes the typical confusions of phonetic units of the unknown speaker and adapt the a-priori probabilities. The learning rates can be dynamically adjusted by the entropy of the a-posteriori probabilities. By that we achieve a very fast adaptation of the a-posteriori probabilities to the optimal recognition rates using a Maximum-Likelihood criterion.

1. INTRODUCTION

Most systems of automatic speech recognition (ASR) achieve very good results even for very large vocabularies, if they are trained and used in a speaker dependent mode. But the recognition results decrease considerably if they are used in speaker independent mode. To close the gap between speaker dependent and speaker independent mode many ASR systems use algorithms to adapt the system to the unknown speaker. Most of these algorithms found in literature try to transform the preprocessed speech signal or to adapt the models of speech used in the classification algorithm. All these algorithms are able to adapt to the typical production of phonetic units and smaller events, including varying dynamics in time and loudness, pitch etc. of the unknown speaker.

Another point is the typical pronunciation of words by the unknown speaker. Most systems of ASR use a phonetical transcription of the words in the lexicon, either to combine phonetic units according to that transcription for recognition ('top down') or to evaluate word hypothesis by comparing the results of the classification with the transcription ('bottom up'). In most cases these transcriptions are drawn automatically from the orthographic representation of the words, eg. by a lookup in a lexicon of pronunciation or by algorithms of speech synthesis. Therefore in most systems of ASR there exists a unique transcription of each word of the lexicon.

Of course only very few speakers (typically well trained wireless announcers) speak a certain word in the way these ASR systems use to represent and therefore expect it to be spoken. In German language we observed the following:

- Some speakers use to pronounce certain phonemes in a special context in a different way, eg. the word 'Berlin' is correctly spoken as (phonetic symbols according to [SAM, 1990]):

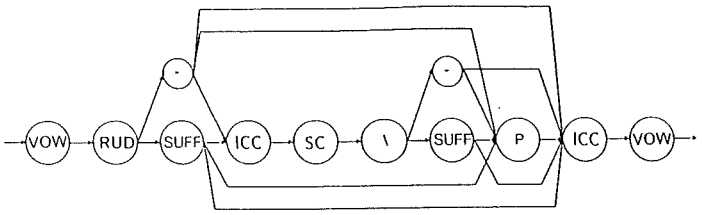

Figure 1: Phonotactic of German language. VOW = vowel, ICC = initial consonant cluster, RUD = rudiment, SUFF = suffix, SC = syllabic consonant, P = silence, -\ = dummy symbols

/b@Rli:n/ (according to [Duden, 1990])

But some speakers (especially those living in Berlin) use to speak 'Berlin' as

/b@6li:n/

Supposing that there is a phonetic unit /R/ used in an ASR system of German language, it has to model the phoneme /6/ in some very special contexts and the phoneme /R/ in all other contexts.

- In fluent German speech most speakers are doing elisions of singles or clusters of phonemes, very often accompanied by certain confusions of phonemes. A good example is the German word 'haben' (to have), which is spoken correctly as

  /ha:b@n/

  Most people especially in the southern part of Germany pronounce 'haben' as

  /ha:m/

  instead. There is a complete elision of the /b/ and the /@/ and additionally the /n/ was transformed into an /m/. Slurring words like this will give serious problems to the ASR system, if it represents the word only according to the correct pronunciation from a lexicon.

We therefore propose a new method of speaker adaptation to achieve pronunciation models typically for the speaker presently using a 'bottom up' system of ASR of German language.

In the following section the ASR system, especially the evaluation of word hypothesis is briefly described. Section 3 is discussing an algorithm to adapt the evaluation of word hypothesis by observing the typical confusions of phoneme clusters in the ASR system. Section 4 is dealing with typical problems, which arise doing an adaptation of pronunciation models. The proposed method of adaption by generalization is described there, too.

2. ASR SYSTEM

The speech signal is preprocessed into a so-called loudness spectrum according to [Zwicker, E., 1982] of dimension 20, the derivation in time of the loudness spectrum and the total loudness every 10 msecs (a concise description of the preprocessing can be found in [Ruske, G. & Beham, M., 1991]). Each of the three features are quantized by a semicontinuous codebook of 265, 128 and 16 symbols respectively as described in [Huang, X.D. & Jack, M.A., 1988]. From that a bottom up classification ([Plannerer, B., 1992]) is done into a chain of consonant clusters and vowels. The classification is increased by using a finite state maschine, which represents the constraints of the German phonotactic according to fig. 1. The definition of the corpus of valid vowels and consonant clusters in German language can be found in [Wolfertstetter, F., 1991]. Additionally the performance is increased by

a pre-segmentation in syllable segments done by an artificial neural network, which constrains the recognition of vowels and syllabic consonants to areas, where there is a high probability of a syllabic nucleus ([Reichl, W., 1992]). The result of the classification is a chain of phonetic symbols of a well defined corpus obeying to the phonotactic mentioned above, eg. the sentence

Aller Dinge Anfang ist schwer.

might lead to the recognition of

/all@Rdi:N@anfaNIstSvER/

(the symbols for dummy and silent segments were omitted for better readability).

The search for the sentence with the highest probability of being spoken is controlled for example by a tree search algorithm as in [Schiel, F., 1991], which evaluates a word hypothesis by comparing the phonetic symbols produced by the classification (see eg. above) with the phonetic symbols belonging to a certain word, eg. 'Dinge' (things), represented in the lexicon as a word model (pronunciation model) like

/dIN@/

Regarding to the fact that both, the result of classification and the lexical entries, obey to the phonotactic mentioned above, the algorithm can very easily determine the a-posteriori probabilities $p(X|Y)$ from the produced phonetic symbol Y to the phonetic symbol X in the word model by a lookup in a squared matrix M, where all the possible a-posteriori probabilities are stored. All these a-posteriori probabilities belonging to a certain word hypothesis are multiplied to achieve the overall probability of that certain word being spoken at a certain position within the speech signal. The tree search algorithm described in [Schiel, F., 1991] uses these probabilities to determine the sentence with the highest probability of being spoken.

Example: The evaluation of the word 'Dinge' (things) at the third syllable position within the classification result above leads to the probability:

$$p(\text{Dinge}, i=3) \quad = \quad p(d|d)\, p(I|i:)\, p(N|N)\, p(@|@)$$

3. ADAPTATION OF THE EVALUATION OF WORD HYPOTHESIS

The idea of this method is very simple: instead of training the matrix M containing the a-posteriori probabilities of all possible confusions once and keeping it constant during the test, we treat the matrix M as being a model of the observed statistics of speaker plus classification. By a supervised or unsupervised backtracking in every recognized sentence we achieve a confusion matrix C by counting the confusions or non-confusions of the consonant clusters and vowels in the sentences regarding to the used word models. After each sentence s_i the confusion matrix C is multiplied with a learning rate $g(s_i)$ and added to the matrix M.

$$M(s_i) = M(s_{i-1}) + g(s_i)\, C(s_i) \tag{1}$$

After that the rows of $M(s_i)$ are normalized to 1 again. The learning rate $g(s_i)$ is either a constant

$$g(s_i) = A_k \tag{2}$$

or calculated for each sentence s_i by

$$g(s_i) = A_k + A_d\, |H_M(s_{i-1}) - H_M(s_{i-2})| \tag{3}$$

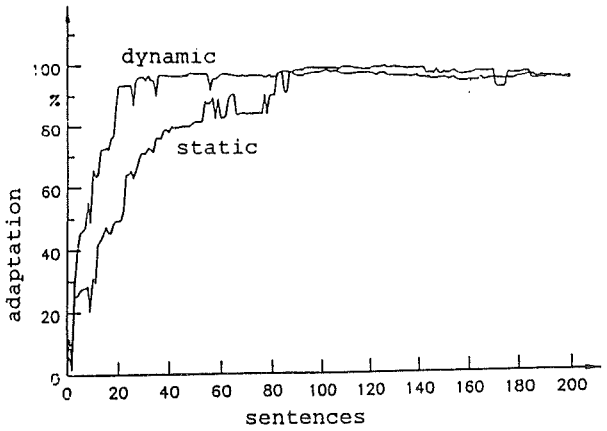

Figure 2: Supervised adaptation over 200 sentences with *static* and *dynamic* learning rates

where H_M is the entropy of the matrix M:

$$H_M = \sum_i \sum_j \left[P(X_i|Y_j)\, ld \frac{1}{P(X_i|Y_j)} \right] \tag{4}$$

and A_d is a constant, which weights the influence of the change of entropy.

That means that the change of entropy within the matrix M containing the a-posteriori probabilities will cause a increase of the learning rate and vice versa. Fig. 2 shows the result of a simulation of a supervised adaption over 200 sentences of 6 words in average. 0 % *adaptation* correspond to the reduced recognition rate caused by a change of speakers. 100 % *adaptation* corresponds to the recognition rate, that can be achieved, if the statistics of the unknown speaker plus classification is exactly known.

As it can be seen, the recognition results labeled *static* according to eqn. 2 achieves convergence after approx. 80 sentences ($A_k = 0.1$).

The curve labeled *dynamic*, where the learning rates were calculated according to eqn. 3, reaches the same convergence after approx. 20 sentences ($A_k = 0.05$, $A_d = 3.0$). We did not found the same good results in unsupervised adaptation. The convergence at 93 % *adaptation* was reached for both methods (static and dynamic learning rates) after approx. 180 sentences.

4. ADAPTATION OF PRONUNCIATION MODELS

Problems

In principle a speaker adaptation of pronunciation models can be done with a statistical approach by observing typical pronunciations like other events in speech. In fact, that approach was investigated at our institute for a ASR system with a reduced amount of 132 words in the lexicon (see [Weigel, W., 1990]). Obviously this method won't work with a lexicon containing more than 2000 words: the time to observe each word at least once would be too long for a realistic speaker adaptation. Therefore we need a way to *generalize* from some observations to as many lexical entries as possible.

We decided to use *rules of differing pronunciation* (RDPs) as a convenient way to generalize from special observations to whole groups of lexical entries.

Example: We observe that a certain speaker uses to speak the word 'gehen' (to go) as

/ge:n/ instead of /ge:h@n/,

which is the correct transcription. We can use this observation to create a RDP like

/h@n#/ ⇒ /n/, which can be read as:
"Substitute the phoneme cluster /h@n/ by /n/, if it occurs at the end of a word (#)"

With that RDP we can automatically transform all lexical entries, which fit to the RDP, eg. the word 'sehen' (to see):

/se:h@n/ is transformed by the rule /h@n#/ ⇒ /n/ into /se:n/

Obviously the next problem is how to create a RDP from very few observations without the supervision of an expert, who can decide whether a RDP is a reasonable generalization or nonsense. We solved this problem by pre-defining a set of 137 RDPs, which cover most the differing pronunciations in German language (without dialects). These RDPs were partly derived from [Jekosch, U. & Becker, T., 1989] and were extended to the use of syllabic consonants at our institute. A complete description of the set of RDPs can be found in [Wolfertstetter, F., 1991].

Algorithm to adapt pronunciation models

First we start with a lexicon containing the standard pronunciation of all words. With the set of 137 pre-defined RDPs we produce a lexicon which contains approx. 3 different possibilities of pronunciation for each word in average. For example: The RDPs

/b@n#/ ⇒ /bm/ and /bm#/ ⇒ /m/

produce from the standard pronunciation of the word 'haben' (to have)

/ha:b@n/ the additional pronunciations /ha:bm/ and /ha:m/.

The produced variations cover the most likely pronunciations in German language free from dialectical peculiarities. Of course the selectivity of the ASR system is reduced by this operation in some cases (approx. 0.1 % of the lexicon will become ambiguous). Now we let the unknown speaker use the ASR system with the expanded lexicon for his purpose. By an unsupervised backtracking of each utterance we determine which types of pronunciation of the words in the lexicon were used by the ASR system to recognize the whole utterance. An index system within the lexicon gives us the information which of the pre-defined RDPs led to this pronunciation. With the statistics of the observed rules we can now adapt the lexicon to the present speaker. For instance we observe the statistics of the RDPs until there is no 'new RDP' (that means a RDP that wasn't observed until now) appearing during an amount of time (eg. 10 sentences). Then we create a new lexicon by using only the prior observed RDPs. By doing that the ASR system can reduce the ambiguity of its lexicon and improves the recognition for the unknown speaker. After a change of speakers there is a need to initialize the lexicon again.

5. CONCLUSION

The proposed algorithms for speaker adaptation were implemented to work in a 'bottom up' ASR system for fluent German language ([Plannerer, B., 1992], [Ruske, G. & Plannerer, B., 1991],

[Hofmann, U., 1991], [Winter, M., 1991]). In preliminary tests with a small amount of data (46 sentences) the adaptation of the evaluation of word hypothesis achieved an improvement from 76.8 % to 78.0 % word recognition in unsupervised adaptation and up to 81.3 % in supervised adaptation. Up to now no further tests were carried out because lack of sufficient data.

Subject to future work will be the collection of data with differing pronunciation in German language from different speakers to evaluate the proposed methods.

REFERENCES

[Duden, 1990] *Band 6, Aussprachewörterbuch*, hrsg. vom Wiss. Rat d. Dudenred.: G. Drosdowski u.a., (Dudenverlag, Mannheim, Wien, Zürich).

[Hofmann, U., 1991] *Automatische Modellierung von Aussprachevarianten in der Spracherkennung*, thesis at the Lehrstuhl für Datenverarbeitung, (Technische Universität München).

[Huang, X.D. & Jack, M.A., 1988] *Hidden Markov Modelling of Speech Based on Semicontinuous Model*, Electronics Letters, Vol. 24, No. 1, pp. 6 - 7.

[Jekosch, U. & Becker, T., 1989] *Maschinelle Generierung von Aussprachevarianten: Perspektiven für Sprachsynthese- und Spracherkennungssysteme*, Informationstechnik it (1989)6, p. 400, (Oldenbourg Verlag).

[Plannerer, B., 1992] *Recognition of Demisyllable Based Units Using Semicontinuous Hidden Markov Models*, Proc. of the International Conference on Acoustics, Speech and Signal Processing, San Francisco, California, pp. 581 - 584.

[Reichl, W., 1992] *Neuronale Netze zur Detektion von Silbenkernen*, Proceedings of the *DAGM Symposium Dresden 1992*, (in print).

[Ruske, G. & Beham, M., 1991] *Gehörbezogene automatische Spracherkennung*, in *Sprachliche Mensch-Maschine-Kommunikation*, pp. 33 - 48, (Oldenbourg-Verlag München Wien).

[Ruske, G. & Plannerer, B., 1991] *Automatische Erkennung fließender deutscher Sprache mit silbenorientierten Einheiten*, in *Studientexte zur Sprachkommunikation*, No. 8, pp. 173 - 182, (Technische Universität Dresden).

[SAM, 1990] *Technical Report SAM*, ESPRIT Projekt 2589, pp. 41 - 43, without place.

[Schiel, F. & Wolfertstetter, F., 1991] *Regelbasierte Erzeugung von robusten Aussprachemodellen und deren Darstellung im Silbenraster*, in *Studientexte zur Sprachkommunikation*, No. 8, pp. 173 - 182, (Technische Universität Dresden).

[Schiel, F., 1991] *Modifizierter A*-Algorithmus zur Erkennung fließend gesprochener Sätze*, in *Informatik Fachberichte 290, Mustererkennung 1991*, pp. 244 - 250, (Springer-Verlag).

[Weigel, W., 1990] *Silbenorientierte Erkennung fließender Sprache mittels diskreter stochastischer Modellierung*, dissertation at the Lehrstuhl für Datenverarbeitung, (Technische Universität München).

[Winter, M., 1991] *Untersuchungen zur Sprecheradaption auf symbolischer Ebene in der automatischen Spracherkennung*, thesis at the Lehrstuhl für Datenverarbeitung, (Technische Universität München).

[Wolfertstetter, F., 1991] *Regelbasierte Generierung und Modellierung von Aussprachevarianten in einem silbenteilorientierten Spracherkennungssystem*, thesis at the Lehrstuhl für Datenverarbeitung, (Technische Universität München).

[Zwicker, E., 1982] *Psychoakustik*, (Springer Verlag, Berlin Heidelberg New York).