DIXIEME COLLOQUE SUR LE TRAITEMENT DU
SIGNAL ET SES APPLICATIONS

NICE du 20 au 24 MAI 1985

A Context-Adaptive Phoneme Lexicon for Continuous Speech Recognition

Otto Schmidbauer

Siemens AG, ZTI INF 112, D-8000 Muenchen 83, Otto Hahn-Ring 6, West Germany

RESUME

La confiance actuelle en "pattern matching" techniques est motiveé par une performance insatisfaisante des ancients systèmes phonétiques de reconnaissance de la parole. La plupart de ces systèmes utilisent une connaisance de la parole relativement petite, en outre ces systèmes evitent explicitement la tâche compliquée de formuler des règles acoustico-phonétiques. Ils déduisent leur performance des générales "pattern matching" methodes. Des experiments récents de la lecture des spectrogrammes ont mené à un intérêt renovelé concernant les approches phonétiques.
Dans ce dossier un lexique phonétique qui respecte des effets reels de coarticulation avec des règles et paramèters acoustico-phonétiques adéquates est présenté. Respectant la tâche délicate de formuler explicitement des règles et paramètres adéquates qui décrivent la variété des réalisations dépendant du contexte des événements acoustico-phonetiques, on propose une methode semi-automatique de la construction et optimisation des regles et parametres.

L'approche présenté est surtout supportée par un set de règles et paramètres qui décrit le comportement des formants dans la domaine de la fréquence et du temps dans des segments stationnaires phonétiques et dans des régions phonétiques de transition. A ce but, le lexique phonétique s'adaptant au contexte pour à peu pres 40 phonèmes de la langue allemande a été développé. L' optimisation des caractères proposés par l'homme et des règles se fait en construisant une parole synthétique, qui respecte des effe ts réels de coarticulation, en utilisant des règles et paramètres du lexique phonétique s'adaptant au contexte, et en comparant cette parole synthétique a une parole naturelle du même contenu. Puis les règles et paramètres sont modifiés d'une monière telle que la distance spectrale entre la parole construite et naturelle est minimisée. C'est la methode d'optimisation des gradients qui arive à la solution de cette tâche délicate.

SUMMARY

Today's reliance on pattern matching techniques is motivated by high recognition rates of such systems, which are not achieved by phonetically based speech recognition systems. Most of these pattern matching systems use relative little speech specific knowledge, moreover, these systems explicitly avoid the complicated task formulating acoustic-phonetic rules. Recent experiments in spectrogramm reading have lead to a renewed interest in phonetically based approaches.
Early phonetically based systems often performed unsatisfactorily,because they didn't consider the context-dependence of acoustic properties of phonemes. In this paper a phoneme lexicon, taking into account real coarticulation effects with adequate acoustic-phonetic rules and parameters for continuous speech recognition is presented. For it is a difficult task of explicitly formulate adequate rules and parameters, which describe the variety of context-dependent realisations of acoustic-phonetic events, a semi-automatic method for generating and optimizing rules and parameters is suggested: Man proposes raw rules and parameters; in an automatic, unsupervised training procedure the rules are modified and the parameters are optimized.
The approach is mainly based on a set of rules and parameters describing formant behaviour in time- and frequency domain. A context-adaptive phoneme lexikon for about 40 phonemes of the german language was developed. The optimization of man-proposed features and rules of the lexicon works by constructing an "artificial" utterance considering real coarticulation effects, using the context-sensitive rules and parameters of the phoneme lexicon, and by comparing this "artificial" utterance with natural speech. Then the rules and parameters are modified in a way that the spectral distance between "constructed" and "natural" utterance becomes a minimum, this delicate task is performed by the optimization procedure "Steepest Descent". Preliminary results show, that coarticulation effects between neighboured phonemes can be modeled by a finite set of adequate rules and parameters.

A Context-Adaptive Phoneme Lexicon for Continuous Speech Recognition

1. Introduction

Despite of intensive research activities in the area of speaker dependent small vocabulary, isolated or connected word recognition, there is still a large gap existing between human and computer recognition of continuous speech : The human speech recognition is robust, versatile and mostly speaker-independent. In comparison, computer recognition systems have to be trained to each new speaker and only perform well, when the test word set is sufficiently acoustically distinctive /2/.

Todays continuous speech recognition systems can mainly be devided into three groups: The Pattern Matching methods, statistic based approaches, and rule and feature based methods.

Today's reliance on pattern matching techniques is motivated by high recognition rates of such systems /7/ and the unsatisfactory performance of early phonetically based speech recognition systems. Pattern matching systems commonly use phonetic units such as words, syllables, demisyllables /6/ or diphones as elementary descision units. A certain degree of speaker-independance is achieved by using the method of multiple templates . Most of these systems utilize relative little speech specific knowledge, besides the assumption, that coarticulation effects occur within an elementary descision unit and that coarticulation effects at unit boundaries are neglectible small /6/. Moreover, these systems explicitly avoid the complicated task of formulating acoustic-phonetic rules; they derive their power from general purpose pattern matching methods.

Statistical pattern recognition approaches are modeling a speech signal as a probabilistic function f.e. of a Markov chain /8/. The fact that short intervalls of speech are similar to more than one phoneme is solved by building a series of Markov models for each vocabulary word. An unknown utterance is recognized by computing a probabilistic score for each possible Markov chain and choose as the recognized token the one with the highest probability.

In opposite to both preceeding approaches the third group of feature and rule based speech recognition systems incorporate extensive knowledge of acoustic-phonetic properties. Recent experiments in spectrogramm reading however suggest that the acoustic signal is richer in phonetic information than previously believed /3,4/. This had lead to a renewed interest in phonetically based approaches to speech recognition with improved understanding of acoustic properties of such sounds in various context-dependent environment /4/.

Rule and feature based systems incorporate the ability to focus attention on the most informative acoustic-phonetic events, whereas systems without phonetic knowledge, as dynamic programming approaches, give equal weight to all parts of a spoken utterance. Because feature based systems evaluate small phonetic differences, it is necessary to extract evident and distinctive features from the speech utterance. The extraction of distinctive phonetic features is commonly not a trivial problem. Therefore sophisticated feature extraction algorithms have to be applied to the speech signal /4,2/.

The unsatisfactory recognition performance of early phonetically based speech recognizers mainly results from the fact, that those systems didn't consider the context-dependence of acoustic properties of phonemes, when spoken in continous speech. At Siemens laboratories in Munich/Germany a knowledge base for a feature and rule based continous speech recognition approach, supported by context sensitive phonemes, is examined /11/. For this purpose a context-adaptive phoneme lexicon for about forty phonemes of the German language was developed.
The rule- and parameter set incorporated in the phoneme lexicon describes the formant behaviour of stationary phoneme parts and phoneme transition regions in time- and frequency domain. The main idea is, that the rules, which decribe the context-sensitive concatenation of phonemes in continous speech, have articulatory background /5/. The context-sensivity considered in this lexicon refers to the acoustic properties of the phonemes preceeding and following the current phoneme.
The delicate task of explicitly formulating adequate rules and parameters, which describe the variety of context-dependent realisations of acoustic-phonetic events, is solved by a semi-automatic method :

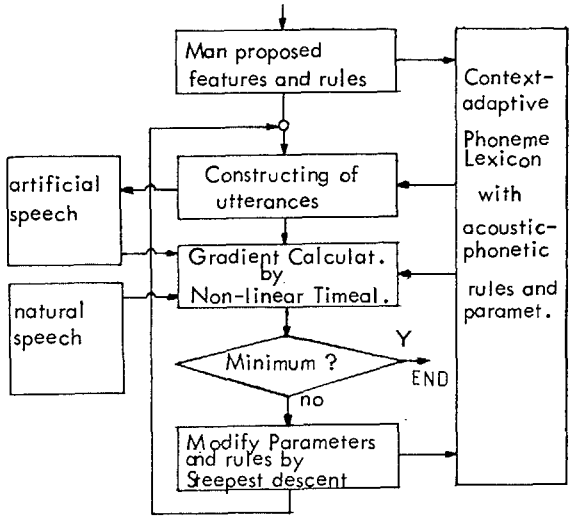

Figure 1: Flowgraph of the total semi-automatic method for optimizing rules and parameters of the context-adaptive phoneme lexicon

Man proposes raw rules and parameters as starting points; during a subsequent non-supervised automatic learning phase, this rule and parameter set is optimized by comparing with natural speech.

The optimization works by constructing an "artificial" utterance considering real coarticulation effects, using the context-sensitive rules and parameters of the phoneme lexicon and by comparing this "artificial" utterance with a natural one of the same contents . This comparison is performed by a non-linear time alignment procedure /1/, which uses severe penalties for time distorsions. Then the rules and parameters are modified in a way, that the spectral distance between "constructed" and "natural" utterance becomes a minimum. This task, working in a multi-dimensional feature space is performed by the optimization procedure "Steepest descent".

The properties of the parameter- and rule set, characterizing the context-dependent acoustic realisations of phonemes, stored in the phoneme lexicon, are defined in chapter 2. The task of concatenating the phonemes with rules and parameters of the lexicon is described in chapter 3. Chapter 4 and 5 deal with the non-linear time alignment procedure, repectively with the optimization procedure. Preliminary results and problems are reported in chapter 6.

2. Basic rules and parameters for context-sensitive phoneme description

Parameters:
- Formant frequencies and bandwidths in target articulation position
- mean duration of static phoneme regions
- duration of phoneme transitions
- factor for correcting static phoneme duration
- centralization factor of formant frequencies and bandwidths of vowels

Rules:
- Interpolation rules for step-wise linear interpolation
- Formant loci for plosives, nasals and fricatives

Table 1: Parameters and rules, which characterize acoustic realisations of phonemes, dependent on manner and place of articulation.

The acoustic features, describing the context adaptive realisation of phonemes in continously spoken speech, have to be devided into two groups :

First, features characterizing stationary phoneme regions, i.e phonemes in target articulation; and second the set of context-dependent rules and parameters describing the great variety of acoustic realisations of phonemes, when spoken in continous speech.

The idea is,that the rules and transition parameters strictly depend on place (f.e. front ,back, ..., labial, alveolar, palatal, ...) and manner (f.e. vowel, fricative, plosive, ...) of articulation of the phonemes to be concatenated, because only articulatory parameters seem to be really speaker - independent. In order to obtain a reasonable starting parameter set for the optimizing procedure, visible speech spectrogramms and formant behaviour in time- and frequency domain of a strictly limited vocabulary (about 30 German words) were examined.

In the first attempt it was assumed, that only consecutive phonemes were coarticulated with each other. Due to the fact that consonants are more coarticulated with neighboured vowels, coarticulation effects between vowels and consonants were characterized more exactly than coarticulation effects between neighboured consonants /6/.

The factor for centralization used in this context means,that the formant frequencies of vowels are moved in the direction of neutral formant frequencies (F1 = .5 kHz, F2 = 1.5 kHz, F3 = 2.5 kHz, ...). Due to the phoneme classes dependent on manner and place of articulation, about 15 to 25 parameters are necessary for a description of one phoneme in arbitrary articulatory context.

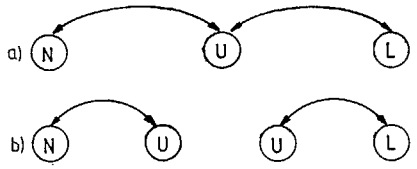

Figure 2: Phoneme Concatenation in the german word zero "/N/ /U/ /L/".
a) presented approach
b) diphon approach

3.Context-sensitive concatenating of phonemes for constructing utterances

Because the context-adaptive lexicon obtains acoustic-phonetic knowledge in terms of rules and parameters, utterances such as syllables, words or sentences may be constructed considering real coarticulation effects between adjacent phonemes. In a further processing step (chapter 4) this utterance is compared with natural speech from different speakers containing identic phonemes, in order to get an idea of the quality of the man-extracted rules and parameters.

A predefined text has to be transformed into a phonetic transcription and then into acoustic parameters. The first transformation is done by a word lexicon with phonetic transcription. The second task is performed by a special purpose algorithm which concatenates consecutive phonemes on feature level, applying the context-adaptive rules and parameters of the phoneme lexicon.This means in praxi an application of step-wise linear interpolation algorithms in the formant, pitch and energy domain, which depend on manner and place of articulation of the phonemes to be concatenated.

Contrary to diphon based systems this approach also takes into account coarticulation effects in stationary phoneme parts (see figure 2,3). That means, if the terminology of the context-adaptive phoneme lexicon is expanded to systems using descision units such as diphones or demisyllables, the phoneme lexicon considers also coarticulation effects between adjacent descision units.

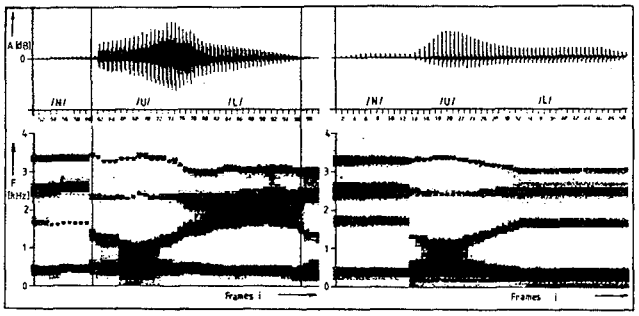

Figure 3: Formant behaviours of an artificially constructed utterance (right) and a natural utterance (left) of the german word "Null". The rule applied here, is a jump in the second formant at the transition from the nasal /N/ to the vowel /U/.

4. Nonlinear time alignment by dynamic programming

The score of a non-linear time alignment procedure between the utterance,constructed by application of knowledge stored in the phoneme lexicon,and a natural utterance describes the likelyhood of both utterances. Because of having different length at least in the beginning of the optimization procedure,it is obvious to use a non-linear comparison method which tolerates time distortions.

The dynamic programming algorithm used here essentially differs from those used in speech recognition applications/1,7/.Severe local penalties for time distortions are introduced to force a linear alignment of natural and "constructed" speech. Furthermore the match is implemented in a way that each time frame of the artificially speech is compared to a time frame of natural speech. That means in terms of speech recognition,the constructed speech is test, the natural speech is reference utterance.

Depending on the three directions horizontal, vertical and diagonal we get the dynamic programming recursion; the accumulated distance D(i,j) in grid point i,j results from the minimum of the sum of old accumulated distance D(i,j) and new local distance d(i,j) :

$$D(i,j) = \min \begin{cases} a * d(i,j) + D(i-1,j\ \ ), \\ b * d(i,j) + D(i,\ \ j-1), \\ \quad d(i,j) + D(i-1,j-1) \end{cases};$$

Distance measures such as the autocorrelation function (ACF) Cityblock and Euklidian distance, the energy-normalized Itakura Saito distance measure and the COSH distance, a symetric issue of the Itakura distance, were examined for that purpose. Latter distances can efficiently be evaluated on the level of autocorrelation and inverse LPC Filter coefficients.
These LPC measures incorporate the advantage, contrary to ACF Cityblock distance, that spectral differences in the regions of formant frequencies are weighted more heavily than spectral differences in dips /9/. Due to that fact LPC distance measures work very well with vowels and sustainable consonants. Spectral distances in voiceless parts of the speech signal are less significant because there are only flat formants and therefore no significant distances. The same is true for the ACF distances in another sense, for this measure accumulates great distances for voiceless speech parts; the total distance of the nonlinear time alignment procedure therefore is not informative.

5.Paramameter and Rule optimization

By minimizing the spectral distance between artificially constructed and natural utterances from alternating speakers the parameters and rules should be adapted to real speech properties.To take into account the context-dependency of the feature set the utterance as a whole has to be optimized (not single parameters of single phonemes).
This results in an optimization problem in a multi-dimensional feature space (in this case up to 200 dimensions!). Therefore the question for converging becomes an important problem. It is advisable optimizing parameters and rules belonging to the time domain with higher priority than spectral ones to aviod a wrong time alignment, i.e. different phonemes are mapped on each

other. Optimizing a rule means to examine the behaviour of the spectral distance dependent on constructing a phoneme transition with or without specific rules.

Besides a modified orthogonal optimization procedure /10/ which diverged sometimes in our tests, the optimization method "Steepest Descent"/10/was applied with good success. Although the converging properties of this method are not optimally, many iteration steps are necessary to reach a minimum, it is converging very sucessfully and securely, even in our multi-dimensional optimization problem.

Using the optimization method "Steepest Descent" the value and the gradient of the function to be minimized are to be evaluated. From the current value the direction of the steepest descent is chosen for the next iteration step. In this case the function to be minimized is the non-linear spectral distance between natural Sn and constructed speech Sc, dependent on the parameter and rule vector $\underline{x}i$ of the phoneme lexicon (i denotes the i-th iteration step) :

$$f(\underline{x}i) = Dis(Sn,Sc)$$

The gradient of the function $f(\underline{x}i)$ is defined by the relation:

$$g(\underline{x}i) = grad(f(\underline{x}i))$$

The direction of the steepest descent is denoted si; it results from:

$$\underline{s}i = -g(\underline{x}i)$$

To go ahead one step c in the direction of si, the new value for the function f in the i-th iteration step is given by:

$$f(\underline{x}i+1) = f(\underline{x}i + c*\underline{s}i)$$

Another problem is the determination of the proper magnitude of c, the value of the basic iteration step. Proceeding from xi in the direction of si the minimum of the function h(c) has to be found :

$$h(c) = min(f(\underline{x}i + c*\underline{s}i))$$

For this problem occurs each iteration step, it has to be solved with a minimum of amount. Initially, c is a relatively large value, to be certain that the minimum to be found is within that intervall. The exact determination is performed by an intervall deviding method called "Goldener Schnitt"/10/.

6. Results

The phoneme lexicon was optimized on the base of twenty german words of two speakers in the first attempt. Preliminary results

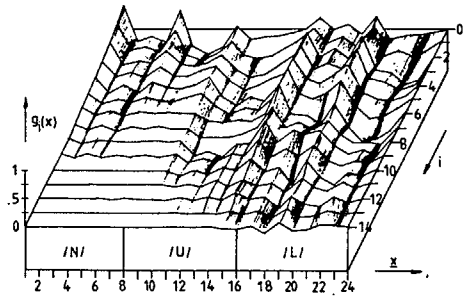

Figure 4: Behaviour of the gradient vector when optimizing the phonemes of the german word zero /N/ /U/ /L/.

were satisfactory by achieving good adaption of the rules and parameters, included in the phoneme lexicon, to real speech. This adaption shows, that real coarticulation effects between any neighboured phonemes can be described and modeled by a set of rules and parameters having articulatory background. By observating and examining gradient vectors during the optimization procedure conclusions about the speaker-independence or -dependence and about the distinctivity of the used rule- and parameter set can be drawn. A further evaluation of the results has to be done in the next future.

Nevertheless, there are some essential problems in tailorihg the appropriate system parameters of the entire optimization system and establishing an adequate starting point in that multi-dimensional feature space, because the problem of local minima during the optimization procedure is likely eliminated, when using a heurisically well defined starting rule and parameter set. In the following we will discuss some of these occuring problems, to get an idea of this complex optimization procedure.

It is obvious that the resulting score of the nonlinear alignment procedure reacts very differently to parameter modifications in frequency-, time-, and energy domain. That fact implies a delicate problem of normalizing parameters of different magnitude and different domains, which has to be solved. Also the score and the the path of the DP is dependent on the applied spectral distance measure and the time-distortion penalties. The last fact worth to be mentioned is to find out an adequate basic iteration step. A large basic iteration step during the gradient calculation implies a fast optimization but also the possibility of over-shooting the absolute minimum; a small iteration step devotes a slow iteration, but incorporates a certain kind of security to reach a real absolute minimum.

This presented optimization systems offers a plausible and highly objective judgement concerning the success or the failure of the optimization by the fact, that artificially speech is constructed and we therefore obtain synthetic speech, after some further transformation; people may judge these constructed utterances by intellegibility and naturaless and not only by a machine calculated scores.

In further work we will try to complete and improve our rule and parameter set to get an exhaustive model for describing coarticulation effects between neighboured phonemes in terms of interpolation rules in the formant-, pitch- and energy domain dependent on place and manner of articulation of the concatenated phonemes. Moreover the the rules and parameters should be examined exhaustively under the the viewpoint of being speaker-dependent or -independent.

Our preliminary results permit the conclusion that the optimization technique "Steepest Descent" is an optimization strategy, which needs a lot of computational amount, but converges very securely. Due to that fact, the computational effort seems to be justificable. This context-adaptive phoneme lexicon is intended to serve as a knowledge base for a future acoustic expert system for a speaker independent speech recognition system.

7. Acknowledgements

This work was sponsored by the german BMFT (Bundesministerium fuer Forschung und Technologie) under the project name SPICOS (Siemens Philips Ipo Continuous Speech recognition).

8. Literature

1. H.D.Hoehne, C.Coker, S.E. Levinson, L.R. Rabiner, "On temporal Alignment of Sentences of Natural and Synthetic Speech", IEEE Tans, Vol. ASSP-26, No.4, Aug.1983, pp. 807 - 813;
2. R.A.Cole, "Performing fine Phonetic Distinctions: Templates vs. Features", Carnegie Mellon University Pittsburgh
3. V.W. Zue, "Implications for Spectrogramm Reading Experiments", Nato ASI, Bonas 1981;
4. V.W. Zue, "Selecting Acoustic Features for Stop Consonants", Proceedings ICASP Boston 1983
5. G.Fant, "Speech Sounds and Features", MIT Press, Cambridge, Massachusetts, London 1973
6. G.Ruske, "On the use of Demisyylables in Automatic Speech recognition", Proc. EURASIP, Erlangen 1983
7. H.Ney, "The use of a one stage Dynamic Programming Algorithm for connected word recognition", IEEE Vol. ASSP-32, 1984
8. S.E.Levinson, L.R. Rabiner, M.M. Sondhi "An Introduction to the Application of the Theory of Probabilistic functions of a Markov Process to Automatic Speech Recognition", Bell System Technical Journal, Vol.62, No.4, Apr.1983
9. A. H. Gray, Jr., "Distance Measures for Speech Processing" IEEE Trans, Vol. ASSP-24, No.5, Oct. 1975;
10. W. Entenmann, "Optimierungsverfahren", Dr. Alfred Huethig Verlag, Heidelberg 1975;
11. O.Schmidbauer, "Optimierung eines Phonemlexikons fuer Spracherkennung durch nicht lineare Anpassung an natuerliche Sprache", Informatik Fachberichte "Mustererkennung", Springer Verlag, Heidelberg, New York 1984;