RELATIONAL PHONETIC FEATURES FOR CONSONANT IDENTIFICATION IN A HYBRID ASR SYSTEM

Jacques Koreman, William J. Barry & Bistra Andreeva

Abstract

In this article we discuss implementation of some fundamental phonetic ideas related to what we shall call "relational processing" in a cross-language consonant identification system. The term relational processing refers to the way vowel transitions play a role in the identification of neighbouring consonants. Two experiments are described: first, consonant identification results from a hidden Markov modelling experiment are presented for consonants plus the preceding and following vowel transitions, if present. The results are compared to a baseline experiment, in which the vowel transitions are not used in the identification of the consonants. In the second experiment, the acoustic parameters are first mapped onto phonetic features; this mapping is performed by a Kohonen network[1]. Since vowel transitions are considered to be particularly important for identification of the place of articulation of the neighbouring consonant, only the place features (and not the consonants' manner features or the phonetic features of the vowel to which the transitions belong) are derived for the vowel transitions. Separate hidden Markov models are trained for the consonants, for the vowel offset and vowel onset transitions which share all consonantal place-of-articulation features. Concatenations of these models form the phone-like recognition units (comparable to the concatenation of phone models for the recognition of words in a conventional ASR system) which are later used for consonant identification. The results are compared with a baseline experiment in which no acoustic-phonetic mapping is performed. The experiments show that relational processing improves the consonant identification results.

In diesem Artikel stellen wir die Implementierung einiger grundlegenden phonetischen Ideen in einem sprachübergreifenden System zur Konsonantenerkennung vor, die unter dem Begriff "relationelle Verarbeitung"

Phonus 3, Institute of Phonetics, University of the Saarland, 1997, 83-109.

fallen. Der Terminus "relationelle Verarbeitung" bezieht sich auf den Beitrag vokalischer Transitionen zur Identifikation benachbarter Konsonanten. Es werden zwei Experimente vorgestellt: zuerst wird in einem Hidden-Markov-Modellierungsexperiment die Konsonantenidentifikationsrate für die konsonantischen Segmente zusammen mit den vorhergehenden und nachfolgenden Vokaltransitionen, falls vorhanden, präsentiert. Die Ergebnisse werden denen eines Basisexperiments gegenübergestellt, in dem die Vokaltransitionen nicht für die Identifikation der Konsonanten verwendet werden. Im zweiten Experiment werden die akustischen Parameter mittels eines Kohonennetzes[1] *zunächst in phonetische Merkmale umgewandelt. Da die Vokaltransitionen insbesondere als "Cue" für die Artikulationsstelle des Nachbarkonsonanten gelten, werden nur die sich darauf beziehenden Merkmale (und nicht die konsonantischen Merkmale, die die Artikulationsart beschreiben, oder die Merkmale des transitionstragenden Vokals) für die Vokaltransitionen definiert. Getrennte Hidden-Markov-Modelle werden für die Konsonanten und für die VK- bzw. KV-Transitionen mit gemeinsamen Merkmalen für die Artikulationsstelle trainiert. Die Konkatenation dieser Modelle bildet die phonartigen Erkennungseinheiten (vergleichbar mit der Konkatenation von lexikalischen Einheiten in der konventionellen ASR), die der späteren Konsonantenerkennung dienen. Die Ergebnisse werden mit dem Basisexperiment verglichen, in dem keine akustisch-phonetische Projektion stattfindet. Die Experimente zeigen, daß die relationelle Verarbeitung die Erkennungsrate erhöht.*

1. Introduction

It has been amply demonstrated that powerful statistical modelling techniques for automatic speech recognition, such as hidden Markov modelling or neural networks (NNs), can lead to very high recognition rates: word recognition rates of more than 95% are no longer exceptional, even for large-vocabulary continuous word recognition systems. Especially, good language models[2] enhance automatic speech recognition (ASR), as has been demonstrated by Lee (1990). In the present article, we shall confine ourselves to the level of phone recognition, though; no

[1] The Kohonen network was developed at the Center for PersonKommunikation at Aalborg University, Denmark. It is described in Dalsgaard (1992).

[2] For different types of language models, see also chapter 8 of Waibel & Lee (1990).

language model is used in the experiments which we shall present. The reason for confining ourselves to the level of phone recognition is that even if language models can solve many of the ambiguities at the phone level, it is important to maximize the phonetic information extracted from the signal. This is especially important as speech technology begins to address the problem of recognizing spontaneous speech (Greenberg, 1996).

Variability in the signal is in part caused by the mutual influence between neighbouring[3] sounds. Because separate states can model these transitions in context-dependent phone models, they result in better recognition rates than context-independent ones (Schwartz et al., 1985). By using generalized triphones, the known acoustic similarities between some transitions are exploited and lead to further improvements in recognition rates (Derouault, 1987; Deng et al, 1988; Lee, 1990). The reason for applying these principles lies especially in the wish to increase homogeneity in the states of hidden Markov models (HMMs) representing vowels, while at the same time maximizing the amount of training data for each HMM. In this article we shall show how using the transitions in a phonetically motivated way also improves consonant identification results.

A set of experiments which focus on relating vowel transitions to the identification of consonants and their place of articulation is presented (see Barry, 1995, for a first discussion of the term "*relational processing*"). All experiments were carried out on 16 kHz labelled speechfiles from the English, German, Italian and Dutch parts of the Eurom.0 database. From the database, the text passages ("numbers passage"), spoken by two female and two male speakers for each language (= 16 speakers), were used. Consonant identification is performed *across* the 4 languages. This stresses the generality of the phonetic principles which are the subject of this article and at the same time increases the training material for HMM.

The consonant identification rates reported in this article are low in comparison to recognition rates presented in the literature. There are several reasons for this, such as the modest amount of training data in the database for some of the HMMs we have trained, the fact that the data were taken from several speakers and from 4

[3] In fact, assimilatory influences need not be due to directly, linearly neighbouring sounds, but can be due to more distant sounds. These longer-distance assimilations are not addressed here.

different languages, the fact that we are only reporting consonant and not vowel identification rates (vowel identification is relatively easy for ASR systems) and the fact that we do not use a language model. We want to stress that for our present purpose, this is not relevant: we have carried out controlled experiments and want to demonstrate that we can improve "conventional" ASR systems by using phonetic strategies. Thus, the actual identification rates should only be considered in a comparison between the experiments and the corresponding baseline experiments.

2. Using vowel transitions to identify consonantal place of articulation

In the introduction, we mentioned the fact that generalized triphones are often used in ASR. The main reason for their success is that the variability in the vowel signal that is being modelled is less than if all vowel transition are modelled in one and the same model. Generalized triphones therefore result in models which are homogenous, without suffering from the disadvantage that there are few data to train the models, as can be the case if vowels are modelled separately for each consonantal context. Interestingly, this stresses the use of vowel transitions for reducing the noise in the signal.

We shall show how vowel transitions, in addition to their role in limiting the noise for the purpose of *vowel* recognition, can play a positive role in the identification of consonants in an ASR system. Since the 1950s and '60s, many experiments have shown the contribution of vowel transitions in consonant identification by human listeners (Cooper et al., 1952; Liberman et al., 1954; Delattre et al., 1955; Delattre, 1968; Stevens & Blumstein, 1978). In most speech recognition systems, though, consonants are modelled only on the basis of their steady states (the part of the speech signal which is traditionally segmented as the consonant). The vowel transitions are only used to model the vowels[4]· We shall present an experiment in which consonants are not only recognized on the basis of their steady

[4] The transitions are implicitly also used for consonant recognition, if the ASR system uses a lexicon, since in the lexicon phone sequences are represented by a concatenation of consonant HMMs and HMMs which model generalized triphones for the vowels.

states, but – if available – also on the basis of the surrounding vowel transitions (see also Cassidy & Harrington, 1995). The results will be compared with a baseline experiment in which the transitions are not used to identify the consonant.

2.1. *Experiment 1*

Preprocessing

Vowel transition were defined to be 35 ms long (compare Furui, 1986). All vowels in the labelfiles were relabelled, dividing the vowel up into a vowel onset transition, a steady state and a vowel offset transition. The duration of the steady state of the vowel was variable; it was computed as the vowel duration in the original labelfile minus 70 ms (which is the total duration of the two vowel transitions). If the vowel in the original labelfile was less than 70 ms, the signal was divided equally into an onset and an offset transition. Since vowel transitions from a particular vowel are similar for all consonants which share the same place of articulation, we generalized across place of articulation. Eight places of articulation were distinguished, namely labial (lab), dental (den), alveolar (alv), alveolo-palatal (alp), palatal (pal), velar (vel), uvular (uvu) and glottal (glo). The transitions were thus labelled as "i:_lab", "O_vel", "alp_u:", etc. (SAMPA notation[5]).

All consonants were represented by the corresponding SAMPA symbols, except for plosives and affricates, which were labelled as a sequence of a closure and friction (if present). The closure was labelled as either voiceless or voiced; labels "p0" and "b0" were used regardless of the consonantal place of articulation, since place of articulation has a negligible effect on the signal. The friction part of the phone was labelled "p, t, k, b, d, g" for plosives and "f, s, S, z, Z" for affricates (according to the phone's voicing and place of articulation). Also, realizations of phones which represent the same phoneme, but are realized very differently, were represented by different labels. This was the case for /r/, which can be realized as a uvular or apical trill ("Ruvu" and "ralv", respectively), or as a retroflex ("rret"), depending on the

[5] SAMPA notation is used as described in the final report of Esprit project 2589: Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation (1992).

language. Similarly, the phoneme /v/ can stand for a labio-dental fricative ("vfri") or approximant ("vapr" = IPA symbol {Ë}), depending on the language[6].

HTK 2.0 was used for parameterization and HMM of the signals (Young et al., 1995). Two versions of the experiment were run, a "static" and a "dynamic" one. In the static experiment, 12 mel-frequency cepstral coefficients (MFCCs) and energy were computed from the signal; in the dynamic experiment, delta coefficients were computed in addition. A 15-ms Hamming window was used, so that the spectral changes in the transitions are not smeared out over time too much; step size was 5 ms. A preemphasis of 0.97 was used. Note, however, that in what we shall term the *dynamic* experiment, static parameters are also used! This is true for both experiments reported in this article.

HMM training and testing

A total of 311 HMMs[7] were trained (HTK programs: HInit and HRest): one for each label in the database[8], except for vocalic steady states. Note that separate HMMs were trained for the transitions and the (parts of) consonants. All the data in the database were used in training as well as in testing (therefore "identification"). The reason for doing so is that the amount of training data available was relatively small. There is no reason to assume that our findings would not hold if training and testing data are different, although this remains to be demonstrated.

In the testing phase (HTK program: HVite), individual consonants, optionally including vowel transitions, were identified, which were all saved as separate testfiles (using the HTK-program HCopy). A total of 12,154 testfiles were thus created.

Besides a file listing all the HMMs, a dictionary and net file are needed for the identification task. In the dictionary, each consonant was defined as a concatenation of a vowel offset transition, consonantal label(s) and a vowel onset transition (comparable to the concatenation of phone models for the recognition of words in a

[6] Of course, phonemes often have different variants even within a language. This is not reflected in the labels, so that we can say that the labelling is phonemic at the level of individual languages and, for most phonemes, across languages.

[7] Each HMM consisted of 3 states with a single Gaussian probability density function (covariance was modelled). The HMMs were left-to-right, no states were allowed to be skipped.

[8] Except for some transition labels which were only represented by one token in the database.

conventional ASR system). The transitions were optional; so were the closure and friction parts of the plosives, with the restriction that minimally the burst had to be present, or the closure plus one transition (since, if there is only a closure label, there is no way of determining the plosive's place of articulation from the label names). All label combinations which can form a consonant are given in the dictionary in appendix A. The dictionary file in appendix A was expanded for this experiment by replacing V in all the transitional labels by all the vowels occurring in such a transition (some of the combinations of a vowel and a consonantal place of articulation did not occur in the database). The expanded dictionary file was used for identification of the consonants.

A very simple net file was used, in which it was stated that each testfile had to be categorized as a consonant, i.e. one of the entries in the dictionary.

Baseline experiment

In parallel to the experiment above, a baseline experiment was carried out. The baseline experiment only differed from the experiment described above in that no vowel transitions were used to identify the consonants. The baseline experiment therefore represents the consonant identification results in a "conventional" HMM recognition system. In the tables, the results for the baseline experiments are given under "no V transition".

2.2. ***Results***

The results show that using vowel transitions improves identification rates. This is true both for the static and for the dynamic experiment, in which MFCCs and energy, and MFCCs, energy and their delta coefficients, respectively, were used. The results are listed in tables 1a and b for identification of the consonant and its place of articulation, respectively.

Table 1a. Identification rates for 34 consonant categories without/with use of vowel transitions, on the basis of static and dynamic input parameters (see text)

	no V transitions	**V transitions**
static	18.78%	19.97%
dynamic	15.04%	17.29%

Table 1b. Identification rates for 8 places of articulation without/with use of vowel transitions, on the basis of static and dynamic input parameters (see text)

	no V transitions	**V transitions**
static	29.20%	43.55%
dynamic	25.90%	41.50%

Despite the fact that the consonants are often not preceded or followed by a vowel transition (the situation that a consonant is flanked by two vowels is the exception rather than the rule in the database), there are considerable differences in the identification rates. This is especially true for the identification of the place of articulation of the consonant. Had we concentrated only on consonants which *do* have transitions preceding and following them, higher identification rates should be expected.

The improvement in the consonant identification rates is only slightly greater in the dynamic experiment. Since delta parameters should better reflect the spectral change in the vowel transitions, this is somewhat surprising.

2.3. *Conclusions*

Table 1a and 1b show that the addition of delta parameters if we want to identify (the place of articulation of) consonants reduces the number of correct identifications (by roughly 2-4%). A possible explanation for the reduction in the identification rates lies in the fact that the consonants themselves (i.e. without the vowel transitions) consist only of steady states, so that the delta parameters do not carry information in the consonants, but instead add noise to the acoustic parameter vectors. We shall return to this in section 6 of this paper, where we discuss the implications for the next phase of our research.

Comparison of table 1a with table 1b shows that the improvement in the correct identification rates for the consonantal place of articulation is quite substantial when we use the vocalic transitions. The improvement in the identification rates of the consonants is small in comparison. The large improvement in the correct place of articulation identification rates confirms the phonetic assumption that the vowel transitions particularly carry information about the place of articulation of the consonant. Thus, it confirms the main phonetic hypothesis underlying the experiment.

If we consider that *all* consonants in the database were offered for recognition and that we did not select only those consonants which are flanked by vowels, the effect is extremely encouraging, since it is only a subset of the total dataset which can have caused the improvement of the identification rates.

3. Recognizing phones on the basis of phonetic features

In the recent literature, several successful attempts have been made to recognize speech on the basis of phonetic features. The phonetic features are either used as input to an ASR system directly or they are used to define the model configurations (Deng and Erler, 1992; Deng & Sun, 1994; Zacks & Thomas, 1994; Bitar & Espy-Wilson, 1995a,b; Strik, 1995; Kirchhoff, 1996a,b). The rationale behind using phonetic features is that they define all and only the relevant distinctions between functional units in speech (phonemes).

In the present experiment, we have used a hybrid ASR system, which combines a Kohonen or self-organizing neural network (SONN) with HMM. The SONN is employed to map acoustic parameters (the static or dynamic acoustic parameters from the first experiment) onto phonetic feature strengths. The phonetic features have been chosen so that they are close to the prime articulatory property determining the acoustic structure. In appendix B, we have listed all the phonetic labels which we used in the present experiment together with their phonetic feature definitions. Features like [±anterior] and [±coronal] or [±grave] and [±acute] are also used in phonological theories, but their relationship to the acoustic realization is much less clear than we may expect for the phonetic features used in the mapping here. The success of the mapping of acoustic parameters onto phonetic features depends on the closeness of the relationship between them. The question of an optimal set of phonetic features for ASR remains an open question, though.

The SONN takes single frames of acoustic parameters as its input and for each frame, outputs a vector of phonetic feature strengths. Since the aim of the experiment is consonant identification, there is no mapping from the acoustic parameters onto vocalic phonetic features, even for the transitions, which are considered to be part of the vowel. The reason is that these features are not relevant for consonant identification. It is important to note that vowel transitions are treated differently from consonants in the acoustic-phonetic mapping. For the consonants, the acoustic parameters are mapped onto the complete set of consonantal phonetic features, i.e. phonetic feature strengths are derived for all consonantal place-of-articulation and manner features. For the transitions, a mapping of the acoustic parameters onto phonetic features is only performed for the place-of-articulation features of the neighbouring consonant, since we want to show the relational function of the transitions, i.e. their importance for the identification of the neighbouring consonant's place of articulation. For transitional frames, the consonantal manner features and all the vocalic features are thus set to zero, since they are assumed to be of minor importance for the identification of the neighbouring consonant. By treating transitions and consonants differently, we help the system to focus on phonetically relevant properties.

In the consonant identification step, the phonetic feature strengths are then used as input to HMM. As in the first experiment, we shall report the results from a consonant identification task.

3.1. *Experiment 2*

Preprocessing

As in the first experiment, consonant and vowel transition labels are used, but this time the vowel transition labels do not only generalize across consonantal place of articulation, but also pool all the vowels, whose identity is not relevant for the identification of the consonant. All the vowels are represented by "V" in the labelname. This produces labelnames such as "V_glo" for the transition of any vowel into a glottal consonant and "den_V" for the transition from a dental consonant into any vowel.

The preprocessing of the signalfiles is exactly the same as in the previous experiment. Again, we shall run two versions of the experiment: one in which only static acoustic parameters are used (MFCCs and energy) and one in which these are used together with dynamic parameters (the MFCC and energy delta parameters).

SONN mapping

A 50x50 SONN was used to map the acoustic parameter vectors (see section 2.1) of all the consonants in the database onto phonetic features. The SONN is described in Dalsgaard (1992). Training of the SONN is performed in two stages:

- In the first stage, called stimulation, all acoustic parameter vectors are fed into the SONN, which organizes itself phonotopically.
- In the second or calibration stage of the training, the same acoustic parameters are fed into the SONN, but this time a phonetic feature vector (the one listed behind the phonetic label in appendix B) is attached to the winning neuron for each frame; which neuron wins is entirely determined by the distance between the acoustic parameter vector and all the neurons in the SONN. All phonetic feature vectors of the frames which activated a neuron during calibration are collected in a matrix for each neuron.

At the end of the training phase, all the phonetic feature values which have activated a neuron are averaged. At this time, all the neurons in the SONN have a vector attached to them which consists of averaged phonetic feature values. If a neuron was often activated by a frame which is belongs to a labial consonant or a vowel transition into or from a labial consonant, the averaged phonetic feature value for [lab] will tend towards 1. If that same neuron was activated non-nasal consonants, the averaged phonetic feature value for [nas] will tend towards -1. In this way, averaged phonetic feature values which potentially range from -1 to 1 are attached to each neuron in the SONN.

The trained SONN is used to map acoustic parameters onto phonetic feature strengths. The mapping is performed as follows: each acoustic parameter vector is fed into the SONN and output phonetic feature strengths are computed by taking a weighted average of the average feature values (see calibration above) of the K-nearest neurons.

HMM training and testing

At the end of the mapping procedure described above, our signalfiles have been converted to files which contain phonetic feature strengths for each frame. The HMM procedure is the same as in the previous experiment, except that the signalfiles contain phonetic feature strengths instead of acoustic parameters and the labelfiles, in which we have not only generalized across the consonants' place of articulation, but also across all vowels, giving only 31 HMMs instead of 311. As in the first experiment, the dictionary containing the consonants to be identified consists of concatenations of HMMs for a vowel offset, (parts of) a consonant and a vowel onset; the dictionary is given in appendix A (no expansion is applied, as in the first experiment; compare section 2.1). The same segments were used for recognition as in the first experiment, except that they were extracted from the phonetic feature strength files instead of from the acoustic parameter files.

Baseline experiment

Experiment 1 of this article forms the baseline for the present experiment (compare tables 1a and 1b above), since no acoustic-phonetic mapping was used and the acoustic parameter vectors were used for HMM directly.

3.2. *Results*

The results for the identification of the consonant and its place of articulation are given in tables 2a and 2b, respectively. Identification rates increase tremendously if we use a hybrid ASR system, in which acoustic parameters are first mapped onto phonetic features, which are then used to train and test HMMs. Consonant identification percentages for the HMM system (conventional HMM on the basis of acoustic parameters, but also using vowel transitions!) are around 25% lower; for place of articulation, the difference is around 15%.

Table 2a. Identification rates with/without acoustic-phonetic mapping on the basis of static and dynamic input parameters (see text) for 34 consonant categories, using vowel transitions

	no mapping	**mapping**
static	19.97%	45.82%
dynamic	17.29%	41.16%

Table 2b. Identification rates with/without acoustic-phonetic mapping on the basis of static and dynamic input parameters (see text) for 8 places of articulation, using vowel transitions

	no mapping	**mapping**
static	43.55%	61.34%
dynamic	41.50%	56.09%

As in the first experiment, performance on the basis of dynamic acoustic parameters (MFCCs, energy and delta coefficients) is lower than if we do not use delta parameters (so-called "static" acoustic parameters). Please note that there is no distinction between the static and dynamic systems as far as the type of input parameters to HMM is concerned: in both cases, only static phonetic features were used (no delta's between the values of the phonetic features in consecutive frames were computed in either the "dynamic" or the "static" system).

3.3. *Conclusions*

In section 3.1 it was explained that we used vowel transition labels in this experiment which generalize across the vowels. This is an additional generalization to the first experiment, in which the vowel transition labels were only generalized across the place of articulation of the consonant. Still, this does not mean that the acoustic variability of the transitions for the different vowels was not respected: at the level of stimulation of the SONNs, acoustic parameter vectors were used to allow the SONNs to self-organize, irrespective of the labels which were later attached to each frame in the calibration stage. I.e., the SONNs can represent the acoustic variability for transitions from or into different vowels by attaching the acoustic parameter vectors to different neurons.

It is in the step to HMM that the generalization becomes relevant. At the level of calibration of the SONN, no vowel features are used, so that the phonetic feature strength vectors which are produced by the acoustic-phonetic mapping no longer reflect the differences between the vowels to which the transitions belong. This was done with a purpose, namely in order to only select information from the transitions which is relevant for the identification of the consonant (or more precisely, its place of articulation).

Performing the same generalization across vowels in the first experiment would not be a useful baseline for the results obtained in the present experiment, since a generalization across vowels at the acoustic level would make the results deteriorate[9]. It would result in inhomogenous HMM reflecting the full range of acoustic variability due to all the different vowels.

The mapping of acoustic parameters onto phonetic feature strengths leads to a large improvement in the identification rates, both for the consonants and for their place of articulation. This shows that the acoustic-phonetic mapping can help the system to focus on linguistically relevant aspects of the acoustically variable realizations, which was the main goal of this experiment.

4. General conclusions

In two experiments, it was demonstrated how phonetic knowledge can be useful in the identification of consonants. Both experiments were concerned with the relational information available in the speech signal. Although the identification rates in the experiments are relatively low, the results are very encouraging. The reasons for the low identification rates were outlined in the introduction and do not affect the interpretation of the results. When we compare the results of our experiments with the corresponding baseline experiments, we can see that the implementation of the phonetic principles addressed here leads to considerable improvements in consonant identification.

In the first experiment, it was shown that we can use the information carried by the vowel transitions to improve the identification of the consonants and particularly of their place of articulation. Since the transitions can be used for the identification of another sound (the consonant) than the one it belongs to (the vowel), this information has been called "relational".

The second experiment was primarily intended to show that we can optimize consonant identification by first mapping acoustic parameters onto phonetic feature strengths, thus highlighting the linguistically relevant information in the signal to

[9] Just for interest's sake, the results if you do use transition labels wich generalize across vowels as well as across consonantal place of articulation when the acoustic parameters are fed into HMM directly are 15.66, resp. 25.66% for consonant identification and identification of its place of articulation on the basis of static acoustic parameters and 12.81 and 21.81% for the corresponding dynamic acoustic parameters.

better distinguish the functional units, which are more abstract and less variable. At the same time, the experiment was concerned with relational aspects of the signal, in that phonetic features representing consonantal place of articulation were extracted from the neighbouring vowel transitions.

5. Data-driven implementations of the phonetic ideas

The experiments and the results which we have presented in this paper show how the implementation of phonetic ideas can improve consonant identification. Since data-driven techniques are often very powerful, it would be good to define the variables used in a data-driven manner.

As an example, we can pick out our definition of the transitions, which we have set at a fixed 35-ms duration. It is clear that transition from vowels into consonants and vice versa do not have a fixed duration. If a transition is shorter than 35 ms, this causes a frame from the steady state of a vowel to be treated as a transitional frame in our system. On the other hand, if a transition is longer than 35 ms, we are only using part of the transition to help to identify the consonant. Both situations can prevent the ASR system from reaching optimal results. It should be possible to select transitional segments from the signal in a data-driven manner, for instance on the basis of the values of the delta parameters or on the basis of a spectral variation function external to the system. First pilots with a spectral variation function led to the insertion of transitional frames in the middle of steady states and also caused parts of transitions not to be selected as such, so that we decided not to proceed along that line. This does not mean, however, that using a data-driven method to select transitions from the signal is impossible; it merely shows that there are no simple solutions to the problem. On the other hand, we should also point out a more principled problem here: some transitions are not characterized by a clear change in the spectrum, so that it may be difficult to identify them as transitions for a system which bases its decision on spectral change. This is the case when the vowel formants and the locus positions of the formants for a consonant are more or less identical, as for instance in combinations of [j] and [i:] or of [w] and [u:]; in these cases the smaller amount of overall spectral energy in the consonant compared to the vowel transitions is the main cue which distinguishes the consonant from the vowel. To a lesser extent, the same problem arises in combinations of for example [d] (locus position of the second formant: 1800 Hz) and [E], where the second formant does not change, while the first formant does; here, too, there is a difference in overall spectral energy between the consonant and the vowel, which can be used as a cue by a spectral variation function..

Another data-driven extension of the ideas presented in this article could be useful for optimizing the set of phonetic features used for acoustic-phonetic mapping (compare experiment 2). Although the feature matrix which we have used in this article has a clear phonetic foundation, it may be far from perfect for the distinction between functional linguistic units on the basis of our parameterization of the acoustic signal (which does not necessarily parallel the representation produced by our hearing system). Automatic clustering techniques may be able to produce a more efficient set of features for this purpose. It is not unlikely that the feature set resulting from automatic clustering, which is optimal from an ASR point of view, is difficult to relate to phonetic or phonological feature sets which have been used by linguists. In a sense, by using automatic clustering to produce a feature matrix for distinguishing all the functional units, we make up for the simplified representation of the speech signal. If we were able to preprocess the speech signal in a way which is more similar to that in human listeners, we should expect the discrepancy between the bottom-up feature set derived by automatic clustering and the more top-down linguistic feature set, which presupposes a human perceptual preprocessor, to be reduced.

We believe there is ample space for data-driven improvements of the ASR strategies suggested in this article. That the strategies we have used can lead to substantial improvements is reflected in the identification rates for consonants and for their place of articulation which were presented in this paper.

6. Future research: using acoustic delta parameters selectively

In the experiments reported, one of the results was that the addition of delta parameters to the static acoustic vector (MFCCs and energy) led to a slight decrease in the consonant identification results. We suggested that this may be due to the fact that the addition of delta parameters for consonants is unnecessary and only presents extra noise or random variation to the ASR system. On the other hand, the delta parameters should be expected to be useful to characterize the vowel transitions, since they supposedly best reflect the spectral change which is typical of most transitions. We have therefore devised a system architecture which makes variable use of the delta parameters, putting a greater weight on the phonetic feature strength vectors output by the dynamic SONN when a frame probably (on the basis of its acoustic parameters) belongs to a vowel transition, while giving more weight to the vector from the static SONN if a frame is more likely to belong to a steady state. It is hoped that using the dynamic information in the acoustic signal selectively will further enhance consonant identification.

7. Acknowledgments

The authors are grateful to Paul Dalsgaard, Ove Anderson, Palli Steingrimsson and Anya V. Hansen, all from the Center for PersonKommunikation (CPK) at Aalborg University (Denmark), for the many cross-disciplinary discussions we have had. We also thank CPK for making the SONN program package available. The ideas which are reported in this article are based on the goals of acoustic-phonetic group in the SPHERE project. Further, they want to thank Andrew Morris for his programming support.

8.. References

Barry, W.J. (1995). *Sphere TMR Network Proposal Description,* section 1.4. Tasks and Assessment Units.

Bitar, N. & Espy-Wilson, C. (1995a). Speech parameterization based on phonetic features: application to speech recognition. *Proc. 4th European Conference on Speech Communication and Technology,* 1411-1414.

Bitar, N. & Espy-Wilson, C. (1995b). A signal representation of speech based on phonetic features. *Proc. 5th Annual Dual-Use Techn. and Applications Conf.,* 310-315.

Cassidy, S & Harrington, J. (1995). The place of articulation distinction in voiced oral stops: evidence from burst spectra and formant transitions. *Phonetica* **52**, 263-284.

Cooper, F., Delattre, P., Liberman, A., Borst, J. & Gerstman L. (1952). Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.* **24**(6), 597-606.

Dalsgaard, P. (1992). Phoneme label alignment using acoustic-phonetic features and Gaussian probability density functions. *Computer Speech and Language* **6**, 303-329.

Delattre, P. (1968). From acoustic cues to distinctive features. *Phonetica* **18**, 198-230.

Delattre, P., Liberman, A. & Cooper, F. (1955). Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.* **27**(4), 769-773.

Deng, L. & Erler, K. (1992). Structural design of hidden Markov model speech recognizer using multivalued phonetic features: comparison with segmental speech units. *J. Acoust. Soc. Am.* **92**(6), 3058-3067.

Deng, L., Lennig, M., Gupta, V. & Mermelstein, P. (1988). Modeling acoustic-phonetic detail in an HMM-based large vocabulary speech recognizer. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* 509-512.

Deng, L. & Sun, D. (1994). A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *J. Acoust. Soc. Am.* **95**(5), 2702-2719.

Derouault, A.-M. (1987). Context-dependent phonetic Markov models for large vocabulary speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* 360-363.

Furui, S. (1986). On the role of spectral transitions for speech preception. *J. Acoust. Soc. Am.* **80**(4), 1016-1025.

Greenberg, S. (1996). Understanding speech understanding: towards a unified theory of speech perception. In: W. Ainsworth & S. Greenberg (eds.). *Proc. Workshop on the Auditory Basis of Speech Perception,* 1-8.

Kirchhoff, K. (1996a). Syllable-level desynchronisation of phonetic features for speech recognition. *Proc. Int. Conf. on Spoken Lang. Proc.,* 2274-2276.

Kirchhoff, K. (1996b). Phonologisch strukturierte HMMs zur automatischen Spracherkennung. In: D. Gibbon (ed.). *Natural Language Processing and Speech Technology (Proceedings of the 3d KONVENS Conference),* 55-63. Berlin: Mouton de Gruyter.

Lee, K.-F. (1990). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transitions on Acoustics, Speech and Signal Processings,* 599-609.

Liberman, A., Delattre, P., Cooper, F. & Gerstman, L. (1954). The role of consonant-vowel transitions in the perception of stop and nasal consonants. *Psychol. Monogr.* **68**(8), 1-13.

Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M. & Makhoul, J. (1985). Context-dependent modeling for acoustic-phonetic recognition of continuous speech. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.*

Stevens, K. & Blumstein, S. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.* **64**(5), 1358-1368.

Strik, H. (1995). Using articulatory knowledge in automatic speech recognition. *Internal Report, Dept. of Language & Speech, University of Nijmegen* (see also: *Newsletter of the Center for Language Studies (CLS)* **9**, 8-13).

Young, S., Jansen, J., Odell, J., Ollason, D. & Woodland, P. (1995). *The HTK Book.* Cambridge: Cambridge University.

Zacks, J. & Thomas, T. (1994). A new neural network for articulatory speech recognition and its application to vowel identification. *Computer Speech and Language* **8**, 189-209.

Appendix A: HTK dictionary file

```
C       hmm_V_pal hmm_C hmm_pal_V
C       hmm_C hmm_pal_V
C       hmm_V_pal hmm_C
C       hmm_C

Dfri    hmm_V_den hmm_Dfri hmm_den_V
Dfri    hmm_Dfri hmm_den_V
Dfri    hmm_V_den hmm_Dfri
Dfri    hmm_Dfri

J       hmm_V_pal hmm_J hmm_pal_V
J       hmm_J hmm_pal_V
J       hmm_V_pal hmm_J
J       hmm_J

L       hmm_V_pal hmm_L hmm_pal_V
L       hmm_L hmm_pal_V
L       hmm_V_pal hmm_L
L       hmm_L

N       hmm_V_vel hmm_N hmm_vel_V
N       hmm_N hmm_vel_V
N       hmm_V_vel hmm_N
N       hmm_N

Ruvu    hmm_V_uvu hmm_Ruvu hmm_uvu_V
Ruvu    hmm_Ruvu hmm_uvu_V
Ruvu    hmm_V_uvu hmm_Ruvu
Ruvu    hmm_Ruvu

S       hmm_V_alp hmm_S hmm_alp_V
S       hmm_S hmm_alp_V
S       hmm_V_alp hmm_S
S       hmm_S

T       hmm_V_den hmm_T hmm_den_V
T       hmm_T hmm_den_V
T       hmm_V_den hmm_T
T       hmm_T
```

```
Z       hmm_V_alp hmm_Z hmm_alp_V
Z       hmm_Z hmm_alp_V
Z       hmm_V_alp hmm_Z
Z       hmm_Z

b       hmm_V_lab hmm_b0 hmm_b hmm_lab_V
b       hmm_b0 hmm_b hmm_lab_V
b       hmm_V_lab hmm_b0 hmm_b
b       hmm_b0 hmm_b
b       hmm_V_lab hmm_b hmm_lab_V
b       hmm_b hmm_lab_V
b       hmm_V_lab hmm_b
b       hmm_b
b       hmm_V_lab hmm_b0 hmm_lab_V
b       hmm_b0 hmm_lab_V
b       hmm_V_lab hmm_b0

d       hmm_V_alv hmm_b0 hmm_d hmm_alv_V
d       hmm_b0 hmm_d hmm_alv_V
d       hmm_V_alv hmm_b0 hmm_d
d       hmm_b0 hmm_d
d       hmm_V_alv hmm_d hmm_alv_V
d       hmm_d hmm_alv_V
d       hmm_V_alv hmm_d
d       hmm_d
d       hmm_V_alv hmm_b0 hmm_alv_V
d       hmm_b0 hmm_alv_V
d       hmm_V_alv hmm_b0

f       hmm_V_lab hmm_f hmm_lab_V
f       hmm_f hmm_lab_V
f       hmm_V_lab hmm_f
f       hmm_f

g       hmm_V_vel hmm_b0 hmm_g hmm_vel_V
g       hmm_b0 hmm_g hmm_vel_V
g       hmm_V_vel hmm_b0 hmm_g
g       hmm_b0 hmm_g
g       hmm_V_vel hmm_g hmm_vel_V
g       hmm_g hmm_vel_V
g       hmm_V_vel hmm_g
g       hmm_g
g       hmm_V_vel hmm_b0 hmm_vel_V
g       hmm_b0 hmm_vel_V
g       hmm_V_vel hmm_b0
```

```
h       hmm_V_glo hmm_h hmm_glo_V
h       hmm_h hmm_glo_V
h       hmm_V_glo hmm_h
h       hmm_h

j       hmm_V_pal hmm_j hmm_pal_V
j       hmm_j hmm_pal_V
j       hmm_V_pal hmm_j
j       hmm_j

k       hmm_V_vel hmm_p0 hmm_k hmm_vel_V
k       hmm_p0 hmm_k hmm_vel_V
k       hmm_V_vel hmm_p0 hmm_k
k       hmm_p0 hmm_k
k       hmm_V_vel hmm_k hmm_vel_V
k       hmm_k hmm_vel_V
k       hmm_V_vel hmm_k
k       hmm_k
k       hmm_V_vel hmm_p0 hmm_vel_V
k       hmm_p0 hmm_vel_V
k       hmm_V_vel hmm_p0

l       hmm_V_alv hmm_l hmm_alv_V
l       hmm_l hmm_alv_V
l       hmm_V_alv hmm_l
l       hmm_l

m       hmm_V_lab hmm_m hmm_lab_V
m       hmm_m hmm_lab_V
m       hmm_V_lab hmm_m
m       hmm_m

n       hmm_V_alv hmm_n hmm_alv_V
n       hmm_n hmm_alv_V
n       hmm_V_alv hmm_n
n       hmm_n

p       hmm_V_lab hmm_p0 hmm_p hmm_lab_V
p       hmm_p0 hmm_p hmm_lab_V
p       hmm_V_lab hmm_p0 hmm_p
p       hmm_p0 hmm_p
p       hmm_V_lab hmm_p hmm_lab_V
p       hmm_p hmm_lab_V
p       hmm_V_lab hmm_p
p       hmm_p
p       hmm_V_lab hmm_p0 hmm_lab_V
```

```
p      hmm_p0 hmm_lab_V
p      hmm_V_lab hmm_p0

ralv   hmm_V_alv hmm_ralv hmm_alv_V
ralv   hmm_ralv hmm_alv_V
ralv   hmm_V_alv hmm_ralv
ralv   hmm_ralv

rret   hmm_V_alv hmm_rret hmm_alv_V
rret   hmm_rret hmm_alv_V
rret   hmm_V_alv hmm_rret
rret   hmm_rret

s      hmm_V_alv hmm_s hmm_alv_V
s      hmm_s hmm_alv_V
s      hmm_V_alv hmm_s
s      hmm_s

t      hmm_V_alv hmm_p0 hmm_t hmm_alv_V
t      hmm_p0 hmm_t hmm_alv_V
t      hmm_V_alv hmm_p0 hmm_t
t      hmm_p0 hmm_t
t      hmm_V_alv hmm_t hmm_alv_V
t      hmm_t hmm_alv_V
t      hmm_V_alv hmm_t
t      hmm_t
t      hmm_V_alv hmm_p0 hmm_alv_V
t      hmm_p0 hmm_alv_V
t      hmm_V_alv hmm_p0

vapr   hmm_V_lab hmm_vapr hmm_lab_V
vapr   hmm_vapr hmm_lab_V
vapr   hmm_V_lab hmm_vapr
vapr   hmm_vapr

vfri   hmm_V_lab hmm_vfri hmm_lab_V
vfri   hmm_vfri hmm_lab_V
vfri   hmm_V_lab hmm_vfri
vfri   hmm_vfri

w      hmm_V_lab hmm_w hmm_lab_V
w      hmm_w hmm_lab_V
w      hmm_V_lab hmm_w
w      hmm_w

x      hmm_V_vel hmm_x hmm_vel_V
x      hmm_x hmm_vel_V
```

```
x       hmm_V_vel hmm_x
x       hmm_x

z       hmm_V_alv hmm_z hmm_alv_V
z       hmm_z hmm_alv_V
z       hmm_V_alv hmm_z
z       hmm_z

p0f     hmm_V_lab hmm_p0 hmm_f hmm_lab_V
p0f     hmm_p0 hmm_f hmm_lab_V
p0f     hmm_V_lab hmm_p0 hmm_f
p0f     hmm_p0 hmm_f

p0s     hmm_V_alv hmm_p0 hmm_s hmm_alv_V
p0s     hmm_p0 hmm_s hmm_alv_V
p0s     hmm_V_alv hmm_p0 hmm_s
p0s     hmm_p0 hmm_s

p0S     hmm_V_alp hmm_p0 hmm_S hmm_alp_V
p0S     hmm_p0 hmm_S hmm_alp_V
p0S     hmm_V_alp hmm_p0 hmm_S
p0S     hmm_p0 hmm_S

b0z     hmm_V_alv hmm_b0 hmm_z hmm_alv_V
b0z     hmm_b0 hmm_z hmm_alv_V
b0z     hmm_V_alv hmm_b0 hmm_z
b0z     hmm_b0 hmm_z

b0Z     hmm_V_alp hmm_b0 hmm_Z hmm_alp_V
b0Z     hmm_b0 hmm_Z hmm_alp_V
b0Z     hmm_V_alp hmm_b0 hmm_Z
b0Z     hmm_b0 hmm_Z
```

Appendix B: Phonetic feature matrix for all consonants

	PLACE							MANNER							S
	lab	den	alv	pal	vel	uvu	glo	plo	fri	nas	lat	apr	tri	voi	ons
p0	0	0	0	0	0	0	0	1	0	0	0	0	0	-1	0
b0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
p	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	0
t	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	0
k	-1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	0
b	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	1	0
d	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	1	0
g	-1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	1	0
f	1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	0
T	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	0
s	-1	-1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	0
S	-1	-1	1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	0
C	-1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	0
x	-1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	0
vfri	1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	0
vapr	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	0
Dfri	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	0
z	-1	-1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	0
Z	-1	-1	1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	0
m	1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	0
n	-1	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	0
J	-1	-1	-1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	0

```
N      -1 -1 -1 -1  1 -1 -1  -1 -1  1 -1 -1 -1  1   0
l      -1 -1  1 -1 -1 -1 -1  -1 -1 -1  1  1 -1  1   0
L      -1 -1 -1  1 -1 -1 -1  -1 -1 -1  1  1 -1  1   0
rret   -1 -1  1 -1 -1 -1 -1  -1 -1 -1 -1  1 -1  1   0
ralv   -1 -1  1 -1 -1 -1 -1  -1 -1 -1 -1 -1  1  1   0
Ruvu   -1 -1 -1 -1 -1  1 -1  -1 -1 -1 -1 -1  1  1   0
j      -1 -1 -1  1 -1 -1 -1  -1 -1 -1 -1  1 -1  1   0
w       1 -1 -1 -1  1 -1 -1  -1 -1 -1 -1  1 -1  1   0
h      -1 -1 -1 -1 -1 -1  1  -1  1 -1 -1 -1 -1 -1   0
V_lab   1 -1 -1 -1 -1 -1 -1   0  0  0  0  0  0  0  -1
V_den  -1  1 -1 -1 -1 -1 -1   0  0  0  0  0  0  0  -1
V_alv  -1 -1  1 -1 -1 -1 -1   0  0  0  0  0  0  0  -1
V_alp  -1 -1  1  1 -1 -1 -1   0  0  0  0  0  0  0  -1
V_pal  -1 -1 -1  1 -1 -1 -1   0  0  0  0  0  0  0  -1
V_vel  -1 -1 -1 -1  1 -1 -1   0  0  0  0  0  0  0  -1
V_uvu  -1 -1 -1 -1 -1  1 -1   0  0  0  0  0  0  0  -1
V_glo  -1 -1 -1 -1 -1 -1  1   0  0  0  0  0  0  0  -1
lab_V   1 -1 -1 -1 -1 -1 -1   0  0  0  0  0  0  0   1
den_V  -1  1 -1 -1 -1 -1 -1   0  0  0  0  0  0  0   1
alv_V  -1 -1  1 -1 -1 -1 -1   0  0  0  0  0  0  0   1
alp_V  -1 -1  1  1 -1 -1 -1   0  0  0  0  0  0  0   1
pal_V  -1 -1 -1  1 -1 -1 -1   0  0  0  0  0  0  0   1
vel_V  -1 -1 -1 -1  1 -1 -1   0  0  0  0  0  0  0   1
uvu_V  -1 -1 -1 -1 -1  1 -1   0  0  0  0  0  0  0   1
glo_V  -1 -1 -1 -1 -1 -1  1   0  0  0  0  0  0  0   1
```

Send to Helmer Strik x, Kathrin Kirchhoff, Johan de Veth x, Entropic (Jon and ?), Aalborg, Alex Strachan, Einar Meister, Andrew x as file Morris, Catia x

Schwartz: page numbers

Read: Schwarz, Derouault, Deng et al. 88 (fit where ref is?).