FIRST EXPERIMENTS ON AN HMM BASED DOUBLE LAYER FRAMEWORK FOR AUTOMATIC CONTINUOUS SPEECH RECOGNITION

Albino Nogueiras, Marta Casar, José A. R. Fonollosa, Mónica Caballero

TALP Research Center
Universitat Politècnica de Catalunya (UPC). Barcelona, Spain.
{albino,mcasar,adrian,monica}@gps.tsc.upc.edu

RESUMEN

The usual approach to automatic continuous speech recognition is what can be called the acoustic-phonetic modelling approach. In this approach, voice is considered to hold two different kinds of information—acoustic and phonetic—. Acoustic information is represented by some kind of feature extraction out of the voice signal, and phonetic information is extracted from the vocabulary of the task by means of a lexicon or some other procedure. The main assumption in this approach is that models can be constructed that capture the correlation existing between both kinds of information.

The main limitation of acoustic-phonetic modelling in speech recognition is its poor treatment of the variability present both in the phonetic level and the acoustic one. In this paper, we propose the use of a slightly modified framework where the usual acoustic-phonetic modelling is divided into two different layers: one closer to the voice signal, and the other closer to the phonetics of the sentence. By doing so we expect an improvement of the modelling accuracy, as well as a better management of acoustic and phonetic variability.

Experiments carried out so far, using a very simplified version of the proposed framework, show a significant improvement in the recognition of a large vocabulary continuous speech task, and represent a promising start point for future research.

1. INTRODUCTION

The acoustic-phonetic approach, mainly using hidden Markov models (HMM), has shown to be a powerful tool for speech recognition [1]. In controlled conditions—i.e., one speaker, using the same recording framework, and in the same favourable environment—recognition accuracies are high, even in very large vocabulary recognition tasks, say automatic dictation. But when conditions degrade and/or change, performance goes down, even for simple tasks as digit strings recognition.

The standard way of dealing with more-than-one speaker, and/or multiple or changing conditions, is to embed all this variability in the model that links the acoustic and the phonetic informations. The framework does not change much whether we expect the speaker to be always the same, using always the same microphone in the same environment, or not. Just the training material changes, incorporating samples of whichever condition we wish our system to be robust.

In this paper we propose the use of a double layer approach to acoustic-phonetic modelling in order to cope—at least partially—with some of the factors that contribute to degrade performance on speaker and recording conditions independence. The main idea is to divide into two layers the standard acoustic modelling: an upper layer, closer to the lexicon; and a lower layer, closer to the acoustic features.

In Section 2, the double layer framework is presented. Section 3 presents the experimentation carried out and the results achieved. Finally, some conclusions and future work are discussed in Section 4.

2. HMM BASED DOUBLE LAYER FRAMEWORK FOR SPEECH RECOGNITION

2.1. Speech Recognition Using HMM's

A standard speech recognition system is based on a set of so called acoustic models that link the observed features of the voice signal with the expected phonetics of the sentence. The most usual implementation of this link is probabilistic, namely HMM's. In this kind of system we can recognise three levels: phonetics, acoustic model, and acoustic features.

The main assumption in standard acoustic phonetic modelling is that phonetic units can be selected such that each word of any vocabulary can be completely expressed by means of them, and that we can estimate the probability density function of these units in the feature space. If this is the case, minimum risk Baye's rule can be effectively used to determine the most probable meaning for a given utterance.

2.1.1. The Phonetics

In the standard approach, it is necessary to establish the expected phonetic transcription of both the training material and the vocabulary to be recognised. In general it can be accepted that languages are quite well represented by means of known phonetic rules. Thus, any task can be expressed in terms of phonetic units that are only dependent on its vocabulary. Nevertheless, it is also well known that there are many exceptions, variants and a wide range of specific problems. For instance, the phonetic transcription rules may change with the speaker's dialect or age group.

Variation at the phonetic level can be treated in several ways. If the speaker is always the same, or we can expect the users of the system to share the same dialectal influences, or we can determine these influences, then we can build a dialect dependent system, with just one transcription per word.

When dialectal dependence is not possible, a common approach is to add alternative pronunciations to the lexicon. But this approach has a big disadvantage: it increases notably the complexity of the grammar, leading to bigger perplexity and higher computational cost. The worst thing here is that, many times, the increased perplexity produces an increase in the error rate that the higher phonetic accuracy does not compensate.

Finally, the third and most usual way of dealing with phonetics variation is to ignore it. In spite of its simplicity, ignoring dialect variations may be enough in many languages and circumstances. The idea of the approach is training with many realisations of each of the possible dialects present in the scope of the task. In this way it is expected that the model embed the different variations providing a single distribution function for all of them. The drawback is straightforward: the acoustic model have to deal with samples of different behaviour, so it becomes more general, less precise. The main advantage is precisely its simplicity. It does not rely in knowing a priori the dialect, does not increase the perplexity of the task, and does not require higher CPU resources.

2.1.2. The Model

An HMM is a collection of *states*. Each frame of voice can be in one and just one of the states at any time. Each HMM is formed of two different parts: a *transition matrix*, and a set of *emission probability functions*. The transition matrix of an N-state HMM is an $N \times N$ matrix. Each element of the matrix represents the probability of moving from one of the states to another. This transition matrix is common to all kind of HMM's, and has shown little influence on the final results.

Each state in the model contains a set of emission probability functions that provide the probability with which this state generates any frame. There are several ways of defining emission probability functions, but the most usual way is by means of mixtures of Gaussian densities, trained with the expectation maximisation algorithm. Yet two alternatives are possible: continuous HMM's, and semi-continuous HMM's.

In continuous HMM's each state is modelled with a mixture of private Gaussians. Gaussians are constructed with diagonal covariance matrix, so independence between components is assumed for each of them. But not for the mixture itself, that will not usually have diagonal covariance. As the number of Gaussians rapidly grows when the number of units and/or states grows, it is usual to tie together groups of them. This means that several states of several units share some of the Gaussian distributions, but not the mixture weights, which are still private.

In semi-continuous HMM's all the Gaussians are shared, and form a vector quantifier. For each frame, the probability density of all the Gaussians is calculated, and a score vector is formed out of the highest of them. The probability of this frame being in a given state is equal to the sum-product of the score vector that represents the frame, and the mixture vector that represents the state. As in the case of continuous models, each Gaussian assumes component independence, but the mixture of them does not. Yet, vector quantisation is rather difficult when the dimension of the vectors grows, so it is usual to assume feature independence. This means, for instance, that the probability of spectrum and energy are calculated at each state separately, and the state probability is given by the product of both.

2.1.3. The Acoustic Features

No matter what structure is used, the probability density functions in the states of the HMM have to model features extracted from the voice signal. The goal of these features will be providing the maximum information about the phonetics of the utterance, while maximally neutralising the effects of the remaining information present in it. Thus, we expect features to be sex and age independent, to be robust in front of noisy or changing conditions, etc. Unfortunately, it is hard to remove all the spurious components, without damaging the needed information.

Features normally used can be grouped in two ways: spectral/energy features, and static/dynamic features. Spectral features are aimed at modelling the spectral envelope, where it is expected to be the maximum phonetic information. Energy is also useful in the characterisation of some sounds and silence. Both spectral and energy features can be static, which means that the feature reflects the behaviour in the current frame; or dynamic, which means that the feature reflects its temporal evolution.

2.2. The Double Layer Framework

In the standard speech recognition framework using HMM's seen above, all the variation in the phonetics and acoustics is put into the same place: the acoustic model. A same model is used for all kind of speakers in all kind of conditions. This produces an intra-unit variance that can be in the same order of magnitude than the variance between units.

The objective of the double layer framework is to reduce the variance of the modelling by separating the acoustic modelling and the phonetic modelling into two linked layers. The lower layer is the acoustic layer. In this layer we pretend to capture all the variability present in the acoustic signal, either it will be considered in the upper layer or not. It has the structure of an acoustic-phonetic classifier, whose results is a score vector formed with the probabilities of each frame being at each state of the classifier. The upper layer is a standard semi-continuous HMM based recogniser, where the feature vectors are no longer probabilities in the feature space, but probabilities in a fuzzy space where the different codewords are assigned to the states of the lower layer classifier.

2.2.1. The Acoustic Layer

In the lower layer of the framework a phonetic-acoustic classifier is required. We call this classifier $\Lambda = \{\lambda_i\}$, where λ_i is the model of the ith unit in the lower layer classifier. At time n, a signal frame $x(n)$ has a probability of being at each of the states of the classifier. Being $q(n)$ the index of the state visited at time n, and q_{ij} the index of the jth state of unit i, we can calculate a score for each frame and state as:

$$\begin{aligned} \mathcal{S}(n,i,j) &= P(x, q(n) = q_{ij}/\Lambda) && (1) \\ &= \alpha_n(x, q_{ij})\beta_n(x, q_{ij}) && (2) \end{aligned}$$

Where $\alpha_n(q_{ij})$ is the standard *forward* probability— the probability of being at state q_{ij} at time n, given $x(k)$ from $k = 1$ to n—, and $\beta_n(q_{ij})$ is the *backward* probability— the probability of arriving from state q_{ij} at time n to the end of the utterance—.

In expression 2 the knowledge of the whole utterance is needed to estimate the score, so it cannot be used in real time applications. One possible alternative is to reduce the scope in which the score is calculated, considering only a certain context of each frame. Defining $x_N(n) = \{x(n-N)\dots x(n)\dots x(n+N)\}$, i.e. the context of x centred at $x(n)$ and of $2N+1$ frames, the score calculation becomes:

$$\begin{aligned} \mathcal{S}(n,i,j) &= P(x_N, q(n) = q_{ij}/\Lambda) && (3) \\ &= \alpha_{N+1}(x_N(n), q_{ij})\beta_{N+1}(x_N(n), q_{ij}) && (4) \end{aligned}$$

The structure of the acoustic models λ_i is free, as long as a score similar to that in expression 4 can be calculated. It is valid for both continuous and semi-continuous hidden Markov models.

2.2.2. The Phonetic Layer

The phonetic layer is a standard semi-continuous HMM framework. An interesting interpretation of semi-continuous models is to consider the vector quantisation step as a part of the feature extraction phase. In this way, speech is no longer represented by its raw features, but by the scores for the Gaussians that form the codebook. Yet, the quantisation step can be also placed in the recognition phase, but separated from the acoustic models. This would be equivalent to consider a double layer framework, where the lower layer is a *blind* classifier: the vector quantisation.

The double layer framework proposed is equivalent to substituting the blind classifier provided by the vector quantisation, with the scores calculated using equation 4. The upper layer must be constructed with semi-continuous HMM's because the new space does not support a reasonable distance definition.

2.3. Expected Advantages of the Double Layer Framework

By itself, the double layer framework is not substantially different from a standard semi-continuous framework. A founded critic is that it does not increase the knowledge given to the upper phonetic layer. It is identical to the standard single layer framework, trained with the maximum likelihood criterion, but not fed with the observation vectors but with other models also trained with the same criterion.

In our opinion, it is true that the double layer framework does not increase the freedom of the system. But information is provided to the system in the form of restrictions. Instead of modelling distribution functions that must embed a lot of variability, making it difficult to estimate them, and making them become wider and less selective, very specialised units are used in the lower level. This specialised units act like a smoothing: once given the probability of being at each of the low level states, the actual position of the frame in the feature space becomes irrelevant.

The strategy is, thus, a may-loose-at-first in order to win-at-the-end one. At first, little gain is expected. From this starting point we plan to add information, otherwise unconsidered, to improve performance. The advantages we expect to get using the double layer framework are of two natures: an increase in robustness, and the capability of modelling information at present not used.

2.3.1. Increased Robustness

We expect to obtain a direct gain in robustness in front of the standard framework because of the above mentioned smoothing effect. Well behaved frames will have a high probability in those states with the highest probability in the upper state. On the contrary, bad behaved frames—due to noise, distortion or bad pronunciation— will have the highest score for the wrong states, forcing the upper level model to accept these confusions as probable during the training phase, but not forcing it to learn the exact position of the erroneous frame.

Besides that, in the lower layer, closest to voice signal, there is no need to use the same task or transcription scheme as in the upper layer. Actually, both the training and the testing material will be re-coded with the results of a classification process, and many of the original phonetic information of the training material will be lost in this re-coding. Thus, in the lower layer it is desirable that the classifier or classifiers used are as accurate as possible, but no matter in which transcription scheme. This can be used in increasing the robustness of the double layer framework in four ways:

◇ Building independent classifiers for each type of feature.
◇ Training the lower layer with the result of the decoding.
◇ Profiting non transcribed material.
◇ Using cross-validation.

By means of using independent classifiers we can get the maximum information from each of the features, while discarding the information not really relevant for distinguishing one sound from another. We may achieve this by using different acoustic units sets adapted to the characteristics of each feature. For instance, it is expectable that different upper-level phonetic units share the same spectrum or energy. If this shared feature is used in building different acoustic models, there is a chance that the wrong decision is made because of marginal differences in the probability distribution functions. But if this sharing is identified in the unit selection process, it can be neutralised by using the same acoustic unit. This procedure may be carried only in the lower layer, not in the upper, where different acoustic units must have different models if any of the features is different. By the same reason, it could neither be used in single layer frameworks.

If we do not really matter what does each classifier decode and, in the end, the results of the decoding will be the only acoustic information driven to the upper layer recogniser, we can increase the accuracy of what the decoder is able to distinguish by training it not with the theoretical phonetic transcription, but with the decoding results of the classifier. Later on this paper it will be experimentally shown that this approach is possible, and it even improves the accuracy of the recognition.

Moreover, if the transcription used in the lower layer is the result of a phonetic decoding, and all the original phonetic information will be discarded, we do not need this information. Thus, non transcribed material may be used to train the lower layer models.

One characteristic of the double layer framework is that both layers may be trained using different training material. In the recognition step, speaker independent decoding will be used to perform speaker independent recognition. In the training step, if the training material is equal in both layers, speaker dependent decoding is used. One way to solve this, and increase the robustness of the framework, is training with different material the upper and the lower layers, performing a sort of cross-validation.

Combining the use of the decoding result as transcription, the use of non transcribed training material, and cross-validation, a not-so-complex training scheme may be proposed. First, initial low level classifiers are trained using the available transcribed data. This initial classifiers are used to provide a transcription to the non-transcribed data. The definitive classifiers are then trained using the non-transcribed data. Finally, the upper layer models are trained with the transcribed data, using the definitive classifiers as lower layer.

2.3.2. Variability Modelling

In the standard single layer framework all the variability modelling effort relays in the acoustic model. Yet, it is well known that there are several variables that, if considered, lead to significant improvement of recognition performance. The most notable of these variables are sex, age, dialect and recording conditions. Incorporating these variables in the single layer framework—or in the upper layer of the herein proposed double layer framework—is very difficult. Instead, including it in the lower layer of the double layer framework is straightforward. We can have, for instance, sex dependent models in the layer closest to the voice signal. In the upper layer, that closest to the phonetics, we may have sex dependent or independent models.

Another source of variability very difficult and expensive to model using single layer frameworks is inter-word contexts. Context is known to have a strong effect on acoustics. Modelling context dependent units leads to notorious improvements over context independent ones, but this modelling is only easy to implement inside the word boundaries. Yet, it is known that context effects are as critical between words as inside them. As in the case of the speaker or environment variables, introducing this information in the lower layer is straightforward, regardless of it being considered or not in the upper layer.

3. EXPERIMENTATION

3.1. The Recognition Task

The task used to test the capabilities of the double layer framework is the recognition of telephonic free speech dialogues in a tourism information retrieval semantic domain. Both the training and the test material used were recorded inside the scope of the European Commission funded project LC-STAR [2], which is devoted to collecting lexica and corpora for automatic speech-to-speech translation.

The corpus is composed of 211 dialogues (422 different speakers), of which 16 define the standard test material. The remaining 195 dialogues are used for training the acoustic models. In mean, each dialogue lasts 9 minutes and is composed of 45 utterances, each of about 30 words. In total, 8418 utterances (29h45m, 230,000 words) were used for training, and 1040 utterances (3h40m, 28,000 words) were used as test material.

The objective of the task is recognising the words pronounced during the tourism information retrieval dialogues. The vocabulary of the task is composed of 7466 words, and the grammar perplexity is around 70. Both vocabulary and grammar were obtained from the training material. The test material presents an out-of-vocabulary rate over 1 %.

3.2. The Baseline System

The system used as baseline is the semi-continuous HMM based system RAMSES [3], using demiphones as acoustic units [4]. The main features of this system are:

- ◇ Speech is windowed every 10ms with 30ms window length. Each frame is parameterised with the first 14 mel-frequency cepstral coefficients (MFCC) and its first and second derivatives, plus the first derivative of the energy.
- ◇ Spectral parameters are quantified to 512 centroids, energy is quantified to 128 centroids. Frames are quantified to 16 centroids for spectral parameters, and to 4 for energy.
- ◇ 1000 demiphones are selected using a minimum entropy decision tree. Left demiphones are modelled with 5 states, right demiphones and silence are modelled with 6 states.
- ◇ State tying is applied following a minimum entropy criterion to reduce the number of distributions from 22,000 (1,000 units × 5.5 states × 4 kinds of feature) to just 2,100.

3.3. The Double Layer Alternatives

Implementing the full double layer framework as described in this paper is a really hard task. At this moment we have just studied some very simplified versions of it in order to assess the utility of the approach.

The main simplification is to reduce the scope of the low level decoding to just on frame, i.e. fixing N in equation (4) to zero. This is a very heavy simplification that reduces greatly the complexity of the system, but will probably have its cost in terms of recognition accuracy.

One implication of reducing the scope to just one frame is that, if the lower level classifiers are feature independent semi-continuous HMM's, the double layer framework is equivalent to a single layer semi-continuous framework using the original features and quantifier. Information added by the double layer will be, if any, in the form of smoothing or restrictions.

The second relevant simplification is that the original transcription using 1,000 demiphones is used for all four kinds of feature, and no information about sex, age, environment or context across words is used.

The five following alternatives have been tested:

Baseline The baseline system with no distribution tying.

Base Tied Baseline with distributions tied to 2,100 vectors.

DL Joint The double layer framework using as classifier the same distributions as used in **Base Tied**.

DL Indep The same framework as **DL Joint**, but training the classifier of each kind of feature independently.

DL Trans The same framework as **DL Indep**, but using as training transcription the phone decoding result.

baseline experiments	Baseline	36.2 %
	Base Tied	38.6 %
double layer frameworks	DL Joint	38.0 %
	DL Indep	37.3 %
	DL Trans	35.9 %

Tabla 1. *Word error rate in the five systems.*

4. CONCLUSIONS AND FUTURE WORK

As it can be readily seen from Table 1, tying the probability density functions produces an increase of almost 2.5 % in the error rate of the baseline. If these same distributions are directly used as classifiers in the double layer framework, recognition improves more than half a point. Further improvement is achieved when the lower layer classifiers are re-trained independently. Finally, when the result from recognising the training material is provided to train the lower layer classifiers, the best result is obtained, even outperforming the untied baseline.

In our opinion, these results, achieved with a very simplified version of what we intend to do in the future, confirm the usefulness of the approach, and encourage us to undertake the whole system. Yet, most of the work has to be done still.

5. BIBLIOGRAFÍA

[1] L, Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] "LC-STAR: Lexicon and corpora for speech to speech translation," IST-2001-32216.

[3] A. Bonafonte et alter, "RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC," in *VIII Jornadas de Telecom I+D*, 1998.

[4] J.B. Mariño et alter, "The demiphone: a new contextual subword unit for continuous speech recognition," *Speech Communication*, vol. 32, no. 3, pp. 187–197, October 2000.