Acoustic-Phonetic Knowledge and Statistics in Automatic Speech Recognition

Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
52056 Aachen, Germany

ney@informatik.rwth-aachen.de

ABSTRACT

This paper deals with the relation of acoustic-phonetic knowledge and its role in automatic speech recognition. Two applications of acoustic-phonetic knowledge are considered in more detail: 1) coarticulation, i.e. the context dependency of the acoustic realizations of phonemes, and 2) speaker adaptation by vocal tract length normalization.

About 25 years ago for each of these two applications, there were already proposals for modelling these types of acoustic-phonetic dependencies. However, in order to successfully exploit these dependencies for automatic speech recognition, the lesson we learned is that a very *careful* integration of these dependencies into the statistical approach is required.

1 Introduction

When we look back at the progress in automatic speech recognition over the last three decades, we observe that there had been plenty of attempts to get away from the statistical approach, such as spectrogram reading, acoustic-phonetic experts systems and artificial intelligence based systems. None of these approaches was found to be competitive with the statistical approach, and the statistical approach became even stronger over the years. For the progress in automatic speech recognition, we can see the following two patterns emerging during the last 30 years:

- Progress is achieved by (more or less) pure statistical methods. Examples are Hidden Markov models, continuous-valued mixture densities, discriminative training criteria and many methods for the adaptation to the speaker and to acoustic conditions in general.
- Progress is achieved by a combination of the statistical and some acoustic-phonetic knowledge. Typically, such a combination goes hand in hand with a suitable training procedure and with the careful exploitation of training data. Examples are the phonetic decision trees for finding generalized context dependent phoneme models and vocal tract length normalization for speaker adaptation.

The organization of this paper is as follows. First, we will consider automatic speech recognition from the statistical point of view and contrast it with the work in phonetics. Then, we will address the use of phonetic knowledge in phonetic decision trees. Finally, we will present the statistical framework of vocal tract length normalization.

2 Automatic Speech Recognition and the Statistical Approach

In building an automatic system for automatic speech recognition, we are faced with the problem of making decisions and learning from examples. It is exactly here where the statistical approach comes in. So far, the most successful approach to automatic speech recognition (ASR) is the statistical one, which is based on the equation:

$$\begin{aligned} \text{ASR} \;=\; & \text{Acoustic–Phonetic–Linguistic Modelling} \\ & + \text{Statistical Decision Theory} \end{aligned}$$

Statistical decision theory provides us with a powerful framework for both making good decisions and estimating parameters from examples.

For the 'low-level' description of speech and image signals, it is widely accepted that the statistical framework allows an efficient coupling between the observations and the models, which is often described by the buzz word 'subsymbolic processing'. The typical advantage in using probability distributions, which also holds for *symbolic* processing in natural language processing tasks, is that they offer an explicit

ISBN 1-876346-48-5

formalism for expressing and combining hypothesis scores:

- The probabilities are directly used as scores. These scores are normalized, which is a desirable property. When increasing the score for a certain element in the set of all hypotheses, there must be one or several other elements whose scores are reduced at the same time.
- It is evident how to combine scores: depending on the task, the probabilities are either multiplied or added.
- Weak and vague dependences can be modelled easily. Especially in spoken and written natural language, there are nuances and shades that require 'grey levels' between 0 and 1.

The most crucial role is taken by the *probability model*, which provides the link between the input data and the output data that have to be produced by the recognition system. The probability model has free parameters that are learned using a suitable *training criterion* from training examples that are representative of the recognition task to be performed. In addition, we have the *decision rule* that is used to select the most suitable output from the many possible outputs.

Even if we think we can manage without statistics, we will need models which always have some free parameters. Then the question is how to train these free parameters. The obvious approach is to adjust these parameters in such a way that we get optimal results in terms of error rates or similar criteria on a representative sample. So we have made a complete cycle and have reached the starting point of the statistical modelling approach again!

When building an automatic system for speech recognition, we should try to use as much prior knowledge as possible about the task under consideration. This knowledge is used to guide the modelling process and to enable improved generalization with respect to unseen data. Therefore in a good statistical modelling approach, we try to identify the common patterns underlying the observations, i.e. to capture dependences between the data in order to avoid the pure 'black box' concept.

When looking at today's speech recognition technology, it is surprising to see how *little* specific phonetic knowledge is used. There are virtually no research groups whose members are *phoneticians*. In the first systems built for automatic speech recognition in the 70's, there seemed to be more *phonetic* groups involved than in today's systems.

The type of acoustic-phonetic knowledge that is used successfully in today's speech recognition systems appears to be rather general and not based on *detailed* phonetic knowledge. Although *phonetics* as such is also concerned with spoken language, there are a number of important *conceptual* differences in the principles:

- In speech recognition, I do not expect hard decisions and simple rules to perform well at the low level. In my view, a negative example is provided by the attempts to come up with what is referred to as *distinctive features*. Why should we expect nature to be so simple?
- Today's high-performance speech recognizers are able to learn from huge amounts of training data, say more than 50 hours of speech. The question is whether similar amounts of data can be processed by a human *phonetician* and whether a consistent and unbiased analysis of the acoustic-phonetic phenomena can be guaranteed. In automatic speech recognition, there are powerful fully automatic training procedures so that subjective effects are ruled out. In addition, there are standard databases so that, at least in principle, other researchers are able to verify the experimental results reported.

3 Context Dependent Acoustic Models and Decision Trees

Phonetic Decision Trees

The problem of automatically defining and learning context dependent acoustic models has been studied for a long time [2, 3, 4, 11]. We describe the variant that is widely used today [14].

After the training of context *independent* phonemes (on a training set of, say, 50 hours of speech), we have a labelled sequence of acoustic vectors $x_1, ..., x_t, ..., x_T$ with *context dependent* states $s_1, ..., s_t, ..., s_T$. To simplify the presentation, we will consider only the Viterbi or maximum approximation and ignore the typical tripartite structure of today's HMMs. For each context dependent HMM state, there is an emission probability distribution $p_\theta(x_t|s_t)$, which is a single Gaussian in the decision tree context.

We consider the triphone contexts in more detail. For a set of 40 phonemes, there is a set of $40^3 = 64\,000$ triphone contexts. This number is too large to reliably train the model parameters for each context. In addition, we face the problem that not all triphone contexts are seen in training, but nevertheless these contexts might come up in the test data. The remedy for both problems is to partition the triphone contexts into equivalence classes:

$$s \rightarrow g(s)$$

using an unknown mapping $g(\cdot)$.

ISBN 1-876346-48-5

To determine the unknown equivalence classes, we resort to statistical estimation and apply the maximum likelihood criterion to the full sequence of acoustic training data:

$$\arg \max_{s \to g(s)} \left\{ \max_{\theta} \left\{ \sum_{t=1}^{T} \log p_\theta\Big(x_t | g(s_t)\Big) \right\} \right\}$$

where the parameters θ of the Gaussian emission distributions are also unknown and must be learned from the acoustic data.

The mapping into equivalence classes is based on a binary decision tree, the leaves of which define the equivalence classes. For each triphone context, we ask binary phonetic questions about the context. Depending on the answer, the right or left child node is selected and so on until we reach a leaf of the tree. The type of phonetic questions is fairly general and refers to the type of predecessor or successor phoneme (e.g. nasal, vowel, fricative, ...). The structure of this phonetic decision tree and the questions are automatically learned from the acoustic training data. Typically [14], by this approach, the word error rate can be reduced by about 10% relative over other approaches to context dependent modelling.

Grapheme-based Acoustic Models

In the approach presented in [6], we go one step further and apply the decision tree concept *directly* to the orthographic spelling of each word to be recognized. The ultimate goal of a speech recognition system anyway is to convert the speech signal into a sequence of written characters. In such a way, we can also handle an unlimited vocabulary, at least in principle.

In the present version, we make the rather crude assumption that one character (or grapheme) corresponds to one acoustic model. The questions in the decision tree are now questions about the predecessor and successor characters. For languages where the pronunciation is close to the orthography (e.g. German, Dutch or Italian), the application of our approach works reasonably well and the performance is comparable with phonetic decision trees. For other languages (e.g. English), obviously, we need an extension so that strings of characters can be considered as a whole, with respect to both the definition of the acoustic models and the questions in the decision tree.

4 Vocal Tract Length Normalization for Speaker Adaptation

The vocal tract length normalization is based on a simple concept: in some approximation, the vocal tract can be considered to be a system of acoustic tubes, and therefore its resonance frequencies are dependent on the vocal tract length [5, 7, 12, 13]. Thus the basic idea for speaker adaptation is to normalize the frequency axis, i.e. to apply a linear scaling of the frequency axis, before the spectral analysis is performed and the pattern recognition process is started. In [9], it is shown that such a scaling of the frequency axis is equivalent to a linear transformation of the acoustic vectors (with constrained matrix entries). The problem now is that the scaling factor itself is unknown and its estimation is not easy, in particular because this has to be done in a fully automatic fashion.

For a quantitative description, we introduce the following notation:

$$\begin{aligned} \text{acoustic vectors:} \quad & X = x_1, ..., x_t, ..., x_T \\ \text{word sequence:} \quad & W = w_1, ..., w_n, ..., w_N \end{aligned}$$

Let us denote the speaker dependent scale factor by α. The class conditional probability $p_\theta(X|W, \alpha)$ is computed using a hidden Markov model framework:

$$p_\theta(X|W,\alpha) = \sum_{s_1^T} \prod_{t=1}^{T} \left[p(s_t|s_{t-1}, W) \cdot p_\theta(x_t|s_t, W, \alpha) \right]$$

with state sequences s_1^T, transition probabilities $p(s_t|s_{t-1}, W)$ and emission probabilities $p_\theta(x_t|s_t, W, \alpha)$. The emission probabilities depend both on some model parameters θ and the unknown scale factor α.

In the statistical spirit [10], we have to integrate out over the unknown random variable α to obtain the α-independent class conditional probability:

$$\begin{aligned} p_\theta(X|W) &= \int d\alpha \; p_\theta(X, \alpha|W) \\ &= \int d\alpha \; p_\theta(\alpha|W) \cdot p_\theta(X|W, \alpha) \\ &\cong \max_{\alpha} \Big\{ p_\theta(\alpha|W) \cdot p_\theta(X|W, \alpha) \Big\} \end{aligned}$$

where, in the last equation, we have applied the widely used maximum approximation.

Thus, under the condition that α is unknown, we obtain the following Bayes decision rule for the unknown word sequence W using a language model $p(W)$:

$$\begin{aligned} &\arg\max_{W} \left\{ p(W) \cdot \int d\alpha \; p_\theta(\alpha|W) \cdot p_\theta(X|W,\alpha) \right\} \cong \\ &\cong \arg\max_{W} \left\{ p(W) \cdot \max_{\alpha} \left\{ p_\theta(\alpha|W) \cdot p_\theta(X|W,\alpha) \right\} \right\} \end{aligned}$$

The same framework can also be applied to linear transformations of the acoustic vectors [8].

To train the unknown model parameters θ, we have to take into account that the scale factor α is unknown for each training speaker. For each speaker r, we are

ISBN 1-876346-48-5

given a pair (X_r, W_r) of acoustic training data X_r and a word sequence W_r.

Using the likelihood criterion, the unknown model parameters θ can be estimated by the optimization [1]:

$$\arg\max_{\theta} \left\{ \prod_{r=1}^{R} \int d\alpha \; p_\theta(\alpha|W_r) \cdot p_\theta(X_r|W_r,\alpha) \right\} \cong$$

$$\cong \arg\max_{\theta} \left\{ \prod_{r=1}^{R} \max_{\alpha} \left\{ p_\theta(\alpha|W_r) \cdot p_\theta(X_r|W_r,\alpha) \right\} \right\}$$

This result is a very complex optimization problem, in particular because there is some interaction between the model parameters θ and the scale factor α, which tends to be masked by the Gaussian mixtures used for the emission probabilities. Methods have been designed to avoid these complications [7, 13].

In the recognition experiments, it is typically found that the vocal tract length normalization reduces the word error rate by 5% to 15% relative over that of the baseline system.

5 Conclusion

Evidently, there is room for improvement in all areas of speech recognition, in particular signal analysis (or feature extraction) and adaptation at all levels: speaking rate, pronunciation lexicon, acoustic conditions (microphone transfer function, room acoustics, ...). The interesting question is how much the phonetic sciences are able to contribute to these future improvements.

REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwarz, J. Makhoul: A Compact Model for Speaker-Adaptive Training. Proc. Int. Conf. on Spoken Language Processing, Philadelphia, PA, Vol. 2, pp. 1137-1140, Oct. 1996.

[2] L. Bahl, P. Souza, P. Gopalakrishnan, D. Nahamoo, M. Picheny: Decision Trees for Phonological Rules in Continuous Speech. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Toronto, Canada, pp. 185-188, May 1991.

[3] P. S. Cohen, R. L. Mercer: The Phonological Component of an Automatic Speech Recognition System. pp. 275-320 in D. R. Reddy (ed.): Speech Recognition. Academic Press, pp. 275-320, 1975.

[4] H. Hon, K. F. Lee: CMU Robust Vocabulary-Independent Speech Recognition System. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Toronto, Canada, pp. 889-892, May 1991.

[5] E. Eide, H. Gish: A Parametric Approach to Vocal Tract Length Normalization. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Atlanta, GA, Vol. 1, pp. 346-349, May 1996.

[6] S. Kanthak, H. Ney: Context-Dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Orlando, FL, pp. 845-848, May 2002.

[7] L. Lee, R. C. Rose: A Frequency Warping Approach to Speaker Normalization. IEEE Trans. on Speech and Audio Processing, Vol. 6, No. 1, pp. 49-60, Jan. 1998.

[8] C. J. Leggetter, P. C. Woodland: Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer, Speech and Language, Vol. 9, No. 2, pp. 171-185, April 1995.

[9] M. Pitz, S. Molau, R. Schlüter, H. Ney: Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. Proc. European Conference on Speech Communication and Technology, Aalborg, Denmark, pp. 2653-2656, Sep. 2001.

[10] A. Sankar, C.-H. Lee: A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition. IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 3, pp. 190-202, May 1996.

[11] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, U. Krasner, J. Makhoul: Context Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Tampa, FL, pp. 1205-1208, March/April 1985.

[12] H. Wakita: Normalization of Vowels by Vocal Tract Length and its Application to Vowel Identification. IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 2, pp. 183-192, April 1977.

[13] L. Welling, H. Ney, S. Kanthak: Speaker Adaptive Modeling by Vocal Tract Normalization. IEEE Trans. on Speech and Audio Processing, Vol. 10, pp. 415-427, Sep. 2002.

[14] S. J. Young, J. J. Odell, P. C. Woodland: Tree-Based State Tying for High Accuracy Acoustic Modelling. Proc. ARPA Human Language Technology Workshop, Plainsboro, NJ, Morgan Kaufmann Publishers, San Mateo, CA, pp. 286-291, March 1994.

ISBN 1-876346-48-5