Dealing with unexpected words

Insperata accidunt magis saepe quam quae speres, i.e. things you do not expect happen more often than things you do expect, warns Plautus (circa 200 BC). Most readers would agree with Plautus that surprising sensory input data could be important since they could represent a new danger or new opportunity. A hypothesized cognitive process involved in the processing of such inputs is illustrated in Figure 1.

In machine recognition, low-probability items are unlikely to be recognized. For example, in the automatic recognition of speech (ASR), the linguistic message in speech data X is coded in a sequence of speech sounds (phonemes) Q. Substrings of phonemes represent words, sequences of words form phrases. A typical ASR attempts to find the linguistic message in the phrase. This process relies heavily on prior knowledge in text-derived language model and pronunciation lexicon. Unexpected lexical items (words) in the phrase are typically replaced by acoustically acceptable in-vocabulary items.[1]

Our laboratory is working on identification and description of low-probability words as a part of the large multinational DIRAC project (Detection and Identification of Rare Audio-Visual Cues), recently awarded by the European Commission. Principles of our approach are briefly described below.

To emulate the cognitive process shown in Figure 1, the contemporary ASR could provide the predictive information stream. Next we need to estimate similar information without the heavy use of prior knowledge. For the estimation of context-constrained and context-unconstrained phoneme posterior probabilities, we have used a continuous digit recognizer based on a hybrid Hidden-Markov-Model Neural-Network (HMM-NN) technique,[1] shown schematically in Figure 2. First, the context-unconstrained phoneme probabilities are estimated. These are subsequently used in the search for the most likely stochastic model of the input utterance. A by-product of this search is a number of context-constrained phoneme probabilities.[2]

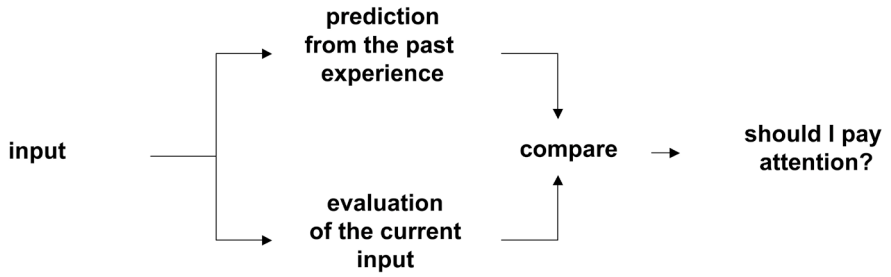

Figure 1. Hypothesized process for the discovery of unexpected items. The sensory input triggers a predictive process in the upper path that uses top-down knowledge from the past experience and generates predicted components of the scene. In parallel, the scene components are also estimated directly (i.e. without the use of the top-down knowledge) from the input. A comparison between the two sets of components may indicate an unexpected item.

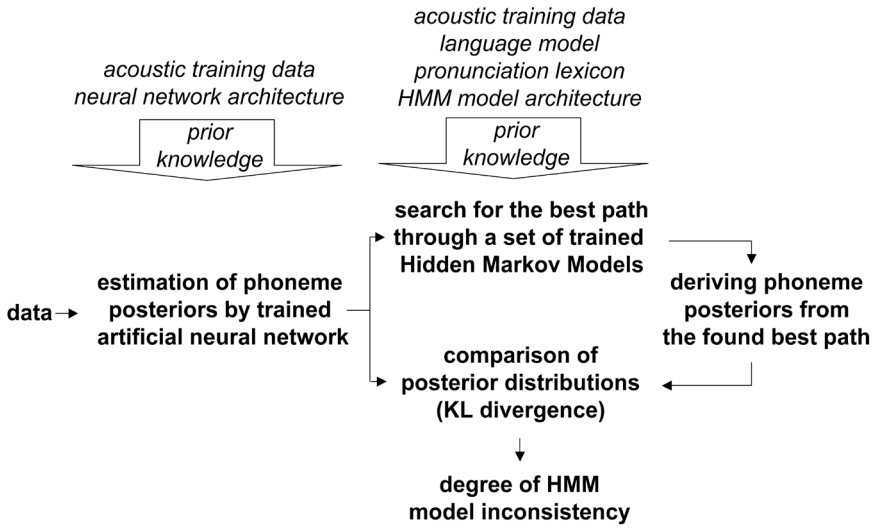

Figure 2. Discovery of out-of-vocabulary words using the hybrid HMM-NN ASR system, in which the out-of-context posterior probabilities estimated by the artificial neural network (ANN) are also directly used in the constrained search for the best model sequence.[2]

The basic principles of deriving the context-unconstrained posterior probabilities of phonemes are illustrated in Figures 3 and 4. A feed-forward artificial neural network is trained on phoneme-labelled speech data and estimates unconstrained posterior probability density function $p_i(Q|X)$.[3] This uses as an input a segment x_i of the data X that carries the local information about the identity of the underlying phoneme at the instant i. This segment is projected on 448 time-spectral basis. As seen in the middle part of Figure 5, the estimate from the NN can be different from the estimate from the context-constrained stream since it is not dependent on the constraints L.

The context-unconstrained phoneme probabilities can be used in a search for the most likely Hidden Markov Model (HMM) sequence that could have produced the given speech phrase. As a side product, the HMM can also yield, for any given instant i of the message, its estimates of posterior probabilities of the hypothesized phonemes $p_i(Q|X,L)$ 'corrected' by a set of constraints L implied by the training-speech data, model architecture, pronunciation lexicon, and the applied language model.[4] When it encounters an unknown item in the phoneme string (e.g. the word 'three' in Figure 5), it assumes it is one of the well known items. Note that these 'in context' posterior probabilities, even when wrong, are estimated with high confidence.

An example of a typical result[4] is shown in Figure 5. As seen in the lower part of the figure, an inconsistency between these two information streams could indicate an unexpected out-of-vocabulary word.

Being able to identify which words are not in the lexicon of the recognizer, and being able to provide an estimate of their pronunciation, may allow for inclusion of

***Hermansky,** continued p. 5*

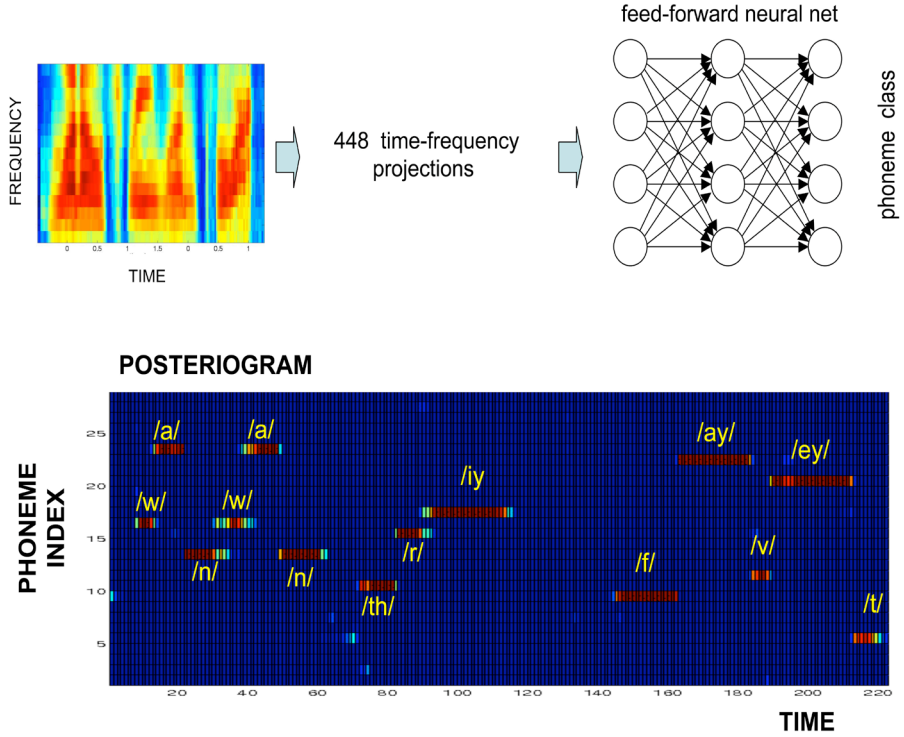

Figure 3. Illustration of the technique for obtaining a reliable estimate of posterior probability density functions $p_i(Q|X)$ without the use of top-down constraints L. The short-term critical-band spectrogram (left part of the figure) is derived by weighted summation of appropriate components of the short-term spectrum of speech. A segment of this spectrogram is projected on 448 different time-frequency bases (shown in Figure 3), centred at the time instant i, yielding a 448 point vector that forms the input to the MLP neural net, trained on about 2 hours of hand-labelled telephone-quality speech to estimate a vector of posterior probabilities $p_i(Q|X)$. A set of $p_i(Q|X)$ for all time instants forms the so-called posteriogram, shown for the utterance one-one-three-five-eight in the lower part of the figure. Higher posterior probabilities are indicated by warmer colors (see Reference 5 for more details).

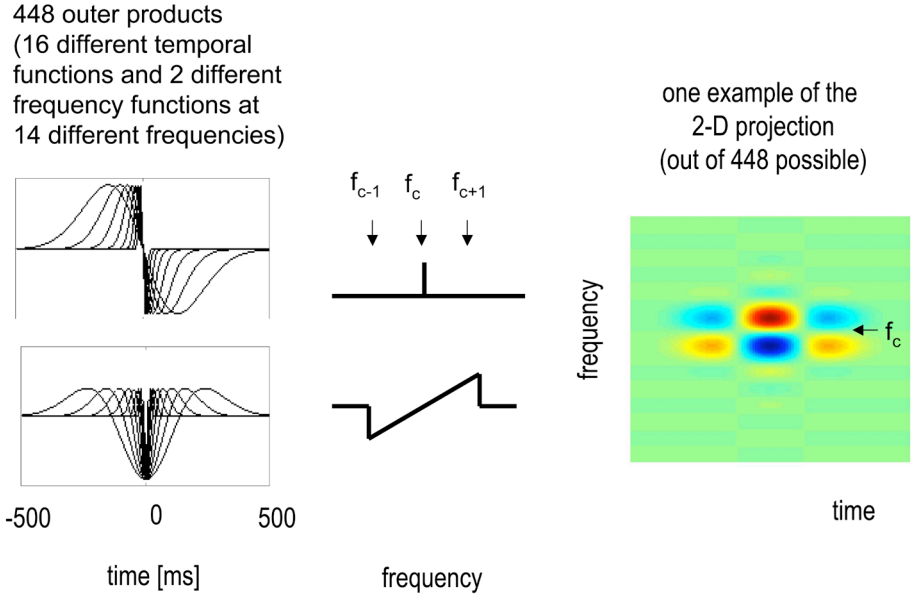

Figure 4. Shown are the time-frequency bases that attempt to emulate some very basic properties of auditory cortical receptive fields (e.g. Shamma). They are formed as outer products of first and second derivatives of truncated Gaussian functions of eight different widths in the time domain, and by summation and differentiation over three frequency components (three critical bands), centred at 14 different frequencies in the frequency domain (see Reference 4 for more details).

Unexpected words, *continued from p. 4*

these new words in the pronunciation dictionary, thus leading to an ASR system that would be able to improve its performance as being it is used over time, i.e. that is able to learn. However, the inconsistency between in-context and out-of-context probability streams need not indicate the presence of unexpected lexical item but could indicate other inadequacies of the model. Further, this inconsistency might also indicate corrupted input data if the in-context probability estimation using the prior *L* yields more reliable estimate than the unconstrained out-of-context stream. Thus, providing a measure of confidence in the estimates from both streams would be desirable when corrupted input is a possibility.

Hynek Hermansky
IDIAP Research Institute
Swiss Federal Institute of Technology
Lausanne, Switzerland
Email: hynek.hermansky@idiap.ch

References

1. H. Hermansky and N. Morgan, *Automatic Speech Recognition,* in **Encyclopedia of Cognitive Science,** L. Nadel, Ed., Nature Publishing Group, Macmilian Publishers, 2002.
2. Bourlard, H. and Morgan, N., **Connectionist Speech Recognition—A Hybrid Approach**, Kluwer Academic Publishers, 1994.
3. H. Bourlard, C. J. Wellekens, *Links between Markov Models and Multilayer Perceptrons,* **IEEE Conf. Neural Information Processing Systems**, 1988, Denver, CO, Ed. D. Touretzky, Morgan-Kaufmann Publishers, pp. 502-510, 1989.
4. H. Ketabdar and H. Hermansky, *Identifying and dealing with unexpected words using in-context and out-of-context posterior phoneme probabilities,* **IDIAP Research Report**, 2006.
5. H. Hermansky and P. Fousek, *Multi-resolution RASTA filtering for TANDEM-based ASR*, in **Proc. Interspeech 2005**, 2005.

Hermansky, *continued p. 10*

Unexpected words, *continued from p. 5*

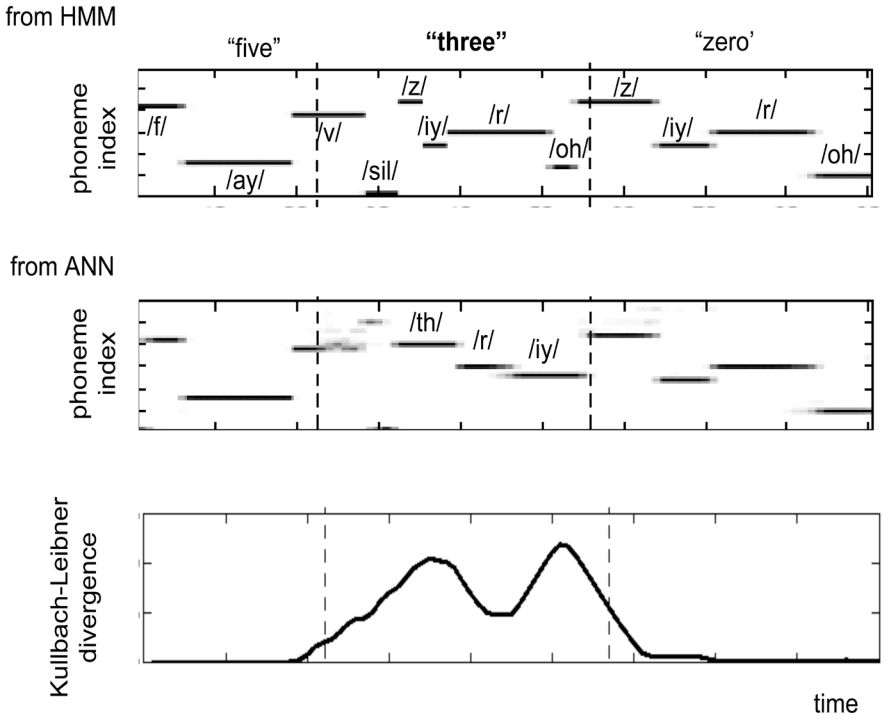

Figure 5. Posterior probabilities of phonemes estimated by an HMM-based system (the upper part of the Figure), and by ANN (the middle part of the Figure). In this example, the HMM model inconsistency was introduced by removing the word three from the recognizer vocabulary. The correct phoneme sequence for the word three is misrepresented in the HMM-derived posteriogram (replaced by a sequence /z/iy//r//oh/ of the in-vocabulary word zero). The ANN-derived probabilities indicate in this case the correct sequence /th//r//iy/ for the out-of-vocabulary word three. Comparison of the respective posterior probability density functions by evaluating their relative entropy (also known as KL divergence): its running average, evaluated over 100 ms time intervals, is shown in the lower part of the figure. This indicates HMM model inconsistency in the neighbourhood of the out-of-vocabulary word there (see Reference 4 on Page 5 for more details).

Spike-based speech, *continued from p. 9*

results are shown in Table 1.

At high signal-to-noise ratio (SNR) values, both systems perform comparably well, but the proposed system using phase synchrony coding is able to outperform the MFCC-HMM algorithm by 12% at 5dB SNR. In regards to the question raised in the title, though applied to a simplified domain, spike-based recognition is clearly more noise robust when compared to a conventional ASR system. This performance is mainly due to the phase synchrony maintaining capabilities of tonotopic neuron populations even under the presence of high amounts of noise.

Future work involves extrapolation of these findings to more complex signals and multi-syllable words by the help of relational networks as observed in the cortex.

Ismail Uysal, Harsha Sathyendra, and John G. Harris
Computational NeuroEngineering Lab
University of Florida
Gainesville, FL, USA
E-mail: ismail@cnel.ufl.edu

References:

1. I. Uysal, H. Sathyendra, and J. G. Harris, *A biologically plausible system approach for noise robust vowel recognition,* **IEEE Proc. of MWSCAS,** CD-ROM, 2006.
2. C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis, *Adaptation in a revised inner-hair cell model,* **J. Acoust. Soc. Am. 113** (2), p. 893-901, 2003.
3. W. Maass, T. Natschlager, and H. Markram, *Real-time computing without stable states: A new framework for neural computation based on perturbations,* **Neural Computation 14** (11), pp. 2531-2560, 2002.