Dependence and independence in automatic speech recognition and synthesis

Simon King

Centre for Speech Technology Research, University of Edinburgh, UK
Simon.King@ed.ac.uk

1 Introduction

When automatically recognising or synthesising speech by computer, we are forced to make a number of assumptions of statistical independence in order to make certain problems tractable. This paper gives a few examples of how phonetic knowledge is already usefully informing these decisions about independence, and a few examples of where it isn't, yet. Temporal integration – how information from a region of speech is related, and is gathered together during perception – is an important aspect of this.

Automatic speech recognition (ASR) and synthesis by computer usually involve the use of various statistical models. For ASR these are models of how acoustic patterns group into larger units (usually phones), how those phones group into words and how words form sentences. For speech synthesis, we need to predict various things, such as: pronunciations for words; durations for phones; where to place phrase breaks and pitch accents; and so on.

In building such models, deciding what factors are dependent and what are independent (in a statistical sense) is crucial. Modelling dependency means a model with more parameters, so the more things we can assume to be independent, the better (so long as those assumptions are close enough to the truth).

1.1 The importance of making statistical independence assumptions

It may not be obvious that we need to make these strong independence assumptions at all. This paper uses two examples (a model of word sequences

and a multivariate Gaussian probability density function) to try to convince you that, more often than not we do not have a large enough corpus of data from which to learn its parameters.

In speech recognition (and other applications) we need to estimate the probability of a word sequence (e.g. a sentence), W where $W = \{w_1, w_2, \ldots, w_L\}$. One way to estimate this would be to count how many times W occurs in a large corpus, and divide by the total number of sentences. That would work, except that the chance of actually finding *any* occurrences of W in a finite-size corpus are very small. We have to *make an independence assumption*, $P(W) \approx \prod_i P(w_i|\text{context of } w_i)$, where the context of w_i which matters is only the identity of the preceding two words. Then, we can estimate $P(w_i|w_{i-1}, w_{i-2})$ by counting how many times $\{w_{i-2}, w_{i-1}, w_i\}$ occurs and dividing it by the number of times $\{w_{i-2}, w_{i-1}\}$ occurs. We are much more likely to find examples of $\{w_{i-2}, w_{i-1}, w_i\}$ than we are to find W. We made an independence assumption – albeit one with only a weak linguistic motivation – which affected the *structure* of the model; model parameters were then learned from from data. Without the independence assumption, the model would have too many parameters to be reliably estimated from a finite-size corpus.

The multivariate Gaussian probability density function is used to model the distribution of observations generated by HMM states. One parameter of this distribution is the square covariance matrix, Σ, containing variances along the diagonal and covariances elsewhere. If the dimension of Σ is large, e.g. 12 Mel-scale cepstral coefficients (MFCCs) + energy + first and second derivatives making a total dimension of 39 by 39, then the number of elements of Σ can quickly become too large to learn from data. If we can assume that there is no covariance between elements of the observation vectors, then Σ becomes diagonal, so the number of parameters is drastically reduced. Making some independence assumptions again results in a model with fewer parameters.

Making the right decision about which factors or model parameters can be assumed to be independent is very important. Phonetics and phonology should inform the model, not the values of its parameters. The way a model represents and uses (in)dependence is part of the *structure* of the model, and not a feature of the particular *values* its parameters take. Therefore, we need to think about ways of altering the model structure itself, and not about how to estimate its parameters. We can almost always get better values for parameters (which generally means parameters that increase the probability of the data, given the model) by estimating them from data than by specifying them using expert knowledge.

2 Some phonetics that mainstream systems already use

Although mainstream speech recognition and synthesis systems use only a small proportion of available phonetic knowledge, there *is* some phonetics in all systems. This section gives a few examples of where using phonetic knowledge in the right way – by using it to make decisions about the type or structure of a model – has made a difference:

2.1 Speech Recognition

Most HMM-based speech recognition systems use context-dependent models of phones, where models of a given phone in *similar* contexts share some parameters. One way to decide which parameters to share is to use a decision tree, with questions about the phonological or phonetic features of the context (Young, Evermann, Kershaw, Moore, Odell, Ollason, Valtchev & Woodland, 2002) . Exactly where in the tree each question goes is determined using data. Phonetic knowledge informs the structure/type of the model (the questions about features), but not the values of its parameters (the order in which the questions are asked).

It is widely accepted, e.g. Wester, Kessens and Strik (2000) , that adding some pronunciation variants to the lexicon of a speech recogniser can improve accuracy, but adding too many variants increases confusability to the point where accuracy goes down. Knowing which words to add variants for and which or how many variants to add requires some phonological knowledge: adding phonologically well-motivated variants can increase accuracy.

2.2 Speech synthesis

Concatenative synthesis is the best method we have for producing synthetic speech. Almost all major commercial synthesisers (e.g. AT&T, Nuance, Rhetorical Systems) use the *unit selection* technique in which units of variable size are selected from a database of several hours of speech, and concatenated to produce the desired output. Synthesis then involves selecting the best possible unit sequence; this is typically done so as to minimise two costs: the *join cost* which measures how well two successive units can be joined, and the *target cost* which measures how closely certain properties (e.g. F_0, duration, stress) of the units from the database match the values predicted by the system. Phonological knowledge is used in the *target cost*, but not currently in the *join cost*.

Tailoring the pronunciation dictionary to the particular speaker is essential for high-quality synthesis, but manually re-writing a dictionary is expensive and time-consuming. One elegant solution to this is the *Keyword Lexicon* (Fitt & Isard, 1999) in which a single underlying lexicon is transformed into an accent-specific one using a relatively simple set of mappings and rules. This is a nice example of phonetic knowledge – Wells' *keyvowel* idea (1982) – informing the model. The phonetic knowledge is used to construct a compact set of "meta-parameters" (the rules and their settings). There are independence assumptions: it is assumed that certain groups of words all use exactly the same vowel (e.g. "bath", "path", ...). A keyword lexicon is downloadable from `http://www.cstr.ed.ac.uk/projects/unisyn/`.

3 Some phonetics that mainstream systems don't use ... yet

3.1 Automatic speech recognition

Long-range dependency (and even local coarticulation) sounds like bad news for ASR - indeed it is for conventional HMMs of phones - but I would argue that it need not be bad news if we see it as an additional source of information. If human speech perception involves temporal integration, then shouldn't automatic speech recognition too?

Using phonological or phonetic features directly for recognition has been proposed and shown to have some potential. These features (high, low, round, etc.) can be detected in speech (King & Taylor, 2000) . Coleman's work (2003), in this volume, demonstrates that evidence can be found in the acoustic signal that correlates with phonological contrasts, but does not show the reverse: that this acoustic evidence can be used to infer the values of phonological features. Detecting features is only a first step towards speech recognition. What is missing is a new type of model that can group features into larger units and account for the observed behaviour of features: asynchrony, spreading, assimilation and so on.

The idea that the speech signal is the result of a number of production mechanisms, with varying degrees of dependence on one another, is attractive for ASR because we can build a factored model that accounts for this. In such a model, speech is modelled as the output of a number of underlying processes (factors), which may optionally be independent of one another. Phonological features are one factorial system, but we can also define factors acoustically (e.g. frequency bands) or infer factors from the data, as in models such as the factorial HMM (Ghahramani & Jordan, 1997) . Whether such models perform better than non-factorial models – e.g. HMMs – remains to be seen.

Hawkins' suggestion that a "speech signal sounds as if it comes from a single talker when it is perceptually coherent, meaning that its properties reflect details of vocal-tract dynamics" (Ogden, Hawkins, House, Huckvale, Local, Carter, Dankovičová & Heid) could include pronunciation consistency within speakers as represented in the phonemic transcriptions of words in the lexicon of a typical ASR system. Although adding pronunciation variation to the lexicon can be beneficial, there is still the problem of confusability. The greatest source of variation is across speakers; within the speech of a single speaker, there is far less variation, so there is a source of information that current systems are not using – pronunciation *consistency* within speakers. So, yes, variants are needed, but *which ones* are needed probably depends mainly on the speaker's accent, and on the styles of speech being recognised. A recogniser which uses this information ought to be more accurate than one that doesn't.

Almost all current ASR systems have models for a manually specified set of phones – that is, a phonetically well-motivated unit. Bacchiani & Ostendorf (1999) showed that a unit inventory can be learned entirely from data. Perhaps a combination of these two approaches would be better than either individually. How to use phonetic or phonological knowledge without going as far as specifying exactly what inventory of units to use is an open question. The choice of unit clearly depends on the acoustic model being used. HMMs and phones appear to be well matched, but for other models phones may not be the best choice.

3.2 Speech synthesis

Current unit selection speech synthesis systems achieve highly intelligible and moderately natural speech. Getting more natural-sounding speech is the next challenge, and this might mean paying more attention to a number of linguistic phenomena, including phonetic detail. For example, whilst perceptually-important fine phonetic detail may be of secondary importance for intelligibility, it may be much more important for naturalness. Without a proven experimental paradigm for measuring naturalness, it's impossible to know for sure. Current evaluation methods for synthetic speech are good at measuring intelligibility, but less good at measuring naturalness; this is an area of ongoing research.

Current systems use purely acoustic (spectral) measures for computing *join cost*, reasoning that if the spectral discontinuity is small, the perceptual prominence of the join will also be small. However, it is becoming apparent from recent work (e.g. Donovan, 2001; Vepa, King & Taylor, 2001) that no single spectral measure works well in all situations. Perhaps we need a measure which pays attention to features with more linguistic basis, like phonological

features or articulation?

No current systems explicitly pay attention to long-term phonetic effects such as long-range coarticulation (West, 1999) or the phonetic detail which Hawkins would argue gives perceptual coherence to the speech signal (Hawkins, 2003, in this volume). We could incorporate this into the *target cost*, but may then need a much larger database of speech to select from in order to find well-matched candidate units.

4 Conclusions

I have argued that, for phonetics to bring real benefits to automatic speech recognition and synthesis, the *way* in which phonetic knowledge is applied is crucial. We need to use such knowledge at a high level. If speech really is the product of numerous underlying process, then we would expect factorial models to be better than HMMs for ASR. If the systematic phonetic detail that gives speech its perceptual coherence aids listeners' comprehension, then synthesis systems that select units with appropriate phonetic detail should achieve higher intelligibility, and presumably naturalness.

References

Bacchiani, M., Ostendorf, M., 1999. Joint lexicon, acoustic unit inventory and model design. Speech Communication 29 (2-4), 99–114.

Coleman, J., 2003. Discovering the acoustic correlates of phonological contrasts. Journal of Phonetics 31, 000–000.

Donovan, R. E., August 2001. A new distance measure for costing spectral discontinuities in concatenative speech synthesisers. In: Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis on CDROM. Atholl Palace Hotel, Perthshire, Scotland.

Fitt, S., Isard, S., Sept. 1999. Synthesis of regional English using a keyword lexicon. In: Proc. Eurospeech 99. Budapest, pp. 823–826.

Ghahramani, Z., Jordan, M. I., 1997. Factorial Hidden Markov Models. Machine Learning 29, 245–273.

Hawkins, S., 2003. Roles and representations of systematic phonetic fine detail in speech understanding. Journal of Phonetics 31, 000–000.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Computer Speech and Language 14 (4), 333–353.

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovicova, J., Heid, S., 2000. Prosynth: an integrated prosodic approach

to device-independent, natural-sounding speech synthesis. Computer Speech and Language 14, 177–210.

Vepa, J., King, S., Taylor, P., 2002. Objective distance measures for spectral discontinuities in concatenative speech synthesis. In: Proc. ICSLP 2002 on CDROM. Denver.

Wells, J., 1982. Accents of English. Cambridge University Press.

West, P., August 1999. The extent of coarticulation: an acoustic and articulatory study. In: Proc. ICPhS 99. San Francisco, pp. 1901–1904.

Wester, M., Kessens, J., Strik, H., 2000. Pronunciation variation in ASR: Which variation to model? In: Proc. ICSLP 2000. Beijing, pp. 488–491.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2002. HTK manual. Cambridge University Engineering Department.