ON ACOUSTIC VERSUS ABSTRACT UNITS OF REPRESENTATION

Daniel Huttenlocher and Meg Withgott

Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, Massachusetts 02139 USA

Stanford University, Center for the Study of Language and Information, Stanford, California 94305 USA

Abstract: Postulating the existence of abstract representational units appears useful in speech research. For instance, such units can be used to partition a large lexicon for word-candidate hypothesization [8] [4], or to specify phonetic deletion and modification sites. However, since such linguistic representations have at best an indirect realization in the physical signal, it has proven difficult to build classifiers for these units. Therefore, recognition systems generally use less abstract units such as spectral templates. We argue that the difficulty of classifying abstract units does not preclude using these units in recognition. In particular, constraint-based systems provide a mechanism for exploiting abstract linguistic knowledge at the acoustic level.

Introduction

Work on lexical and phonological representation assumes the existence of abstract units such as phonemes or allophones. Powerful general principles have been identified operating under this assumption. However, attempts at developing recognizers which use similar units have met with difficulty (cf. [6]). Thus, systems for classifying the acoustic signal generally use representations which are far less abstract (e.g., templates, vector quantized spectra, etc.).

We consider some of the reasons that it is difficult to recognize abstract units such as phonemes from the speech signal. Then we turn to the limitations of current recognition systems. Finally we suggest how some of these limitations may be overcome by formulating lexical and phonological knowledge as constraints on acoustic data.

These constraint-based models can be used to specify that certain acoustic patterns are consistent with a given word. They may also specify that certain acoustic information is inconsistent with the presence of a given word. The critical idea is that of viewing recognition as consistency checking. This idea contrasts strongly with the use of abstract units in transformational systems.

Recognizing Abstract Units is Hard

The difficulty of recognizing abstract units such as phones or diphones from the speech signal is attributable to several factors. First is the problem of segmenting the speech signal into phonetic-sized units. Certain regions of an utterance do not clearly correspond to any particular phoneme or other abstract unit. Furthermore, segmentation errors cause the insertion and deletion of phonetic units.

Second is the difficulty of classifying the segments that have been identified. Variation across talkers causes a given abstract unit to have different realizations for different talkers. These may even overlap, as in the case of /s/ and /ʃ/. Phonetic sized units can also be difficult to classify because they are distorted due to contextual effects (e.g., the /t/ in a /tr/ cluster). Third, certain regions of an utterance are often difficult to classify, such as unstressed syllables.

Thus, a given classifier will perform very well only in certain regions of an utterance, or for certain talkers. This suggests letting the classifier do "only as much as can be done reliably." However, this means that no single abstract level of representation is sufficient.

On top of all this, having identified a sequence of abstract units it is still difficult to do word recognition. Part of the problem is the phonological variation in the production of individual words. Deletion, epenthesis, and other phonological modifications can cause extreme departures from the canonical form.

The problem of mapping from a sound sequence to words is even harder in the case of continuous speech because the limit of the match is not generally known. For instance, it is well known that in fluent speech the phrase "did you go to the.." (/dɪd#yuʷ#goʷ#tʰuʷ#ðə/) can be produced as [dɪʃəgəɪɔəðə].

Considerable attention has been paid to the problem of recognizing words from phonetic sequences. The most common approach is to formulate transformational rules which characterize phonological variation. Such rules map lexical baseforms to surface phonetic strings. This mapping is then either used to expand each lexical entry into all possible surface forms, or to transform an input sequence into its possible underlying forms [7]. However, this assumes that all pronunciations can be anticipated and captured by the rules. Furthermore, since these rules are based on phonetic transcriptions, it is assumed that the output of the classifier is adequately detailed and relatively error free. These assumptions have not been borne out in actual speech.

Current Recognition Systems are Limited

Using acoustic representations for recognition seemingly bypasses the problems of classification and retrieving the underlying phonemic form. However, such systems only work for restricted tasks. While the IBM recognizer [1] is perhaps the most successful system to date, it appears to be reaching the limit of the approach.

The IBM recognizer searches the entire lexicon in recognizing each word. The most obvious consequence of this is the large amount of computation required. A more serious problem is that the distance between an unknown word and each lexical entry does not provide very strong discrimination among the possibilities. This is partly due to the fact that distance metrics are sensitive to acoustic differences, whereas phonological processes can cause large acoustic differences between pronunciations of the same word. These differences can be as large as those between different words, as when "balloon" is pronounced "b'loon", which is acoustically similar to "bloom".

As a result, the IBM system relies heavily on word tri-gram probabilities for its performance. These probabilities are obtained by observing word triples in a large training corpus. However, the use of tri-gram models makes it difficult to add new words because their probabilites must be estimated. Furthermore, tri-grams are not good models of novel sentences even from the same vocabulary. For a 1.8 million word corpus of text, the tri-grams found in one 1.5 million word subset covered only 77% of the tri-grams observed in the remaining 300,000 words [5].

Thus while tri-grams provide substantial constraint, they are too specific in that they don't capture general properties of English. However, a more general characterization of allowable word sequences is unlikely to provide nearly as much constraint. For example, attempts at using syntactic constraints in speech recognition have

required using artificially simple grammars to appreciably limit the possible word candidates [6]. Therefore, some other source of constraint will be needed in order to develop the next generation of recognition systems.

A Look at Using Abstract Units in Recognition

There are three potential advantages of using abstract representational units in recognition. First, exploiting phonological information as a source of constraint in recognition requires using an abstract representation. Second, training a system (or adapting it to new speakers) can be greatly simplified by the use of abstract units. Third, abstract representations enable the use of non-exhaustive matching techniques in lexical access.

With respect to the problem of training, abstract sound units can be used to bootstrap the training process by representing each word in terms of component parts. Training then operates over this smaller set of units rather than over words. In a very large vocabulary system, such a bootstrapping process appears necessary. For example, the IBM system uses phonetic-sized units for training.

With respect to the problem of matching and lexical access, there are two ways in which abstract units can be used. The first is to search only part of the lexicon, rather than matching against every entry and picking the best match. The second is to match against only some of the information in each lexical entry being considered, depending for example on the certainty of the classifier.

While the use of abstract units can theoretically address such issues, the fact of the matter is that systems have been relatively unsuccessful at using abstract units. We claim that this can be traced to the framework within which abstract properties have been formulated, rather than to the use of abstract units per se.

For instance, if phonological rules captured the variability in speech, then lexical access could simply be done by table lookup. Yet as we noted above, there is substantial variability which cannot be accounted for by rules, and this causes classification errors. Thus, the transformational formulation does not get around the problem of exhaustive search of the lexicon.

Another approach which uses abstract units is to characterize what is stable or reliable about a given lexical entry, rather than trying to capture variability. This approach has been taken by Shipman, Zue and Huttenlocher in their work on partitioning the lexicon into equivalence classes of words sharing the same features. For example, manner of articulation features can be used to partition a 20,000 word lexicon into classes of only about 30 words on average.

Using this approach, ideally only that subset of the lexicon corresponding to a given feature sequence must be searched in lexical access. However, this assumes that each word has a small number of partial representations as output by the classifier. While the proposed partial representations are less sensitive to variability than phonetic representations, this still may not be a reasonable assumption.

Conclusion

In the previous section we have seen that systems which use abstract phonetic units have been developed based on the assumption that these units have reliable acoustic correlates. One example of this was transformational systems which view recognition as mapping between sequences of abstract units. In order to apply these transformations, the abstract units must first be reliably classifiable from the acoustic signal. Abstract units often do not have reliable acoustic manifestations, however. The absence of these correlates has led to the development of acoustically-based systems which do not use linguistic constraints at all.

While abstract units do not have reliable acoustic correlates, a given abstract unit is only consistent with certain acoustic patterns. Since constraint-based models can be used to specify what acoustic information is consistent with a given abstract unit, they are a convenient formalism for expressing such knowledge. In particular these models provide a means for expressing partial and redundant information [9] [2] [3]. This ability to exploit multiple levels of specificity means the classifier can be allowed to do as much as it can, while still using a lexical partitioning based on abstract representational properties.

Acknowledgments

This work was supported in part by the Defense Advanced Research Projects Agency under Office of Naval Research Contracts N0014-82-K-0727 and N0014-80-C-0505 to M.I.T., in part by Schlumberger Computer Aided Systems, and in part through an award from the System Development Foundation to the Center for the Study of Language and Information at Stanford.

References

1. Bahl, L.R., A.G. Cole, F. Jelinek, R.L. Mercer, A. Nadas, D. Nahamo, and M.A. Picheny "Recognition of Isolated Word Sentences from a 5000-Word Vocabulary Office Correspondence Task", *Proc. IEEE ICASSP*, 1983.

2. Fenstad, J. E., P-Kr. Halvorsen, T. Langholm and J. van Benthem. (To appear) *Equations, Schemata and Situations: A Framework for Linguistic Semantics* Dordrecht: Reidel. Also in CSLI-TR-29, Stanford Univ., 1985.

3. Grimson, W.E.L. and T. Lozano-Perez "Recognition and Localization of Overlapping Parts from Sparse Data", MIT Artificial Intelligence Laboratory Memo No. 841, 1985.

4. Huttenlocher, D.P. "Exploiting Sequential Phonetic Constraints in Recognizing Words", MIT Artificial Intelligence Laboratory Memo No. 867, 1985.

5. Jelinek, F. "Problems of Language Modeling for Speech Recognition", in preparation, IBM T.J. Watson Res. Ctr.

6. Klatt, D. Review of the ARPA Speech Understanding Project, *J. Acoust. Soc. Am.*, Vol. 62, No. 6, December 1977.

7. Oshika, B. T., V. W. Zue, R. V. Weeks, H. Neu, and J. Aurbach. The Role of Phonological Rules in Speech Understanding Research, *IEEE Trans. on Acoustics, Speech, and Signal Processing.* Vol. ASSP-23, 1, February 1975.

8. Shipman, D. and V.W. Zue "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems" *Proc. IEEE ICASSP*, 1982.

9. Sussman, G.J. and G.L. Steele "CONSTRAINTS: A Language for Expressing Almost Hierarchical Descriptions" *Artificial Intelligence*, Vol. 14, No. 1.