LEXICAL REDUNDANCY IN SPEECH PERCEPTION: VIGILANT MEMORY

THOMAS R. HOFMANN

Beginning from the introduction of the sound spectrograph, many years have been spent discovering the acoustic correlates or language-significant sounds, discovering the means for extracting these significant units. There has also been the basic question, that is, what sort of linguistic units allow successful extraction: phonons (or phonetic features), phonemes, syllables, or some larger units?

Much has been learned and much improvement has attended the extraction task: however, I am convinced that further improvement will come mostly with better technology for faster and simpler extraction, and not in significantly lower error rates, either by choosing a unit of a different size or by discovery of acoustic features which are more reliable indicators. I base this on two facts: people simply do not pronounce ideally, and people persist in talking in noisy environments, where some of the information needed for identification is lost.

Nevertheless, much vigorous effort continues: a good portion of this Congress is related in one way or another to this aim. In a sense, the present paper stands outside of this interest, because its aim is not to determine acoustic correlates of some units and how to extract them from the signal, but to determine what to do with them after they are extracted.

But, of course, what is to be done with the extracted features—i.e., the purpose of their extraction—should influence the questions posed concerning extraction. If in no other way, it should define what remains to be done, and determine when further improvement is useless. I shall not attempt such evaluation here, as I think it must wait on general acceptance of a basis for judgement.

The facts mentioned above, that pronunciation is often imprecise[1] and that background noise sometimes overrides the signal, may be treated together as noise (Interdialectal interference may also be included). This sort of noise can degrade a language signal to the point where the human being cannot determine it correctly without

[1] Imprecise in the sense that in running speech, as opposed to carefully enunciated speech, goals of articulatory behaviour are often not attained, metatheses and spoonerisms occur, and occasionally a segment other than the one intended by the speaker is uttered.

context. This entails the obvious: that the speech perception mechanism uses higher-level redundancies to correct errors from their context.

It is clear that there is an effective method of using these redundancies to reconstruct the intended signal. If the intended signal could not be effectively reconstituted in spite of such noise, human beings would pronounce more accurately, and would choose less noisy environments for speech (or speak more loudly and design better communication systems).

It is also clear that, if we are to build a device to extract the linguistic signal out of an acoustic wave, it will be useful to know what these redundancies are and how to take advantage of them to reconstruct the signal. This knowledge is of course also necessary for a model of speech perception.

What are these redundancies which account for context-dependent improvement of recognition? There are the following possibilities.

First, starting from the smallest, we note that not all combinations of phonetic features exist in a language; the set of phonemes is smaller than it could be. Redundancy of this type, however, could not account for improved human performance with additional context.

Second, of the phonemes in a language, not all possible sequences are found in the language. The set of syllables used in a language is sometimes smaller than the possibilities thereof—and there are also systematic restrictions on the possibilities. This type of redundancy cannot account for the usefulness of context beyond the syllable and hence is not a candidate for the higher level redundancy we seek (though I suspect that the systematic constraints have not been fully exploited for reducing errors).

As a third possibility, we may suggest syntactic phrases as a basis for using the inherent redundancies for error correction. However, syntax relates only parts of speech or lexical items. Before syntactic constraints can be applied to detect an error, the lexical classes or parts of speech must be assigned to the portions of the input. Hence the recognition of lexical items must come before the use of syntactic redundancy.

A fourth possibility lies in the use of semantic restriction by the context. Bruce (1956) has shown an enhanced recognition ability where the listener expects words chosen from a semantically-constrained vocabulary such as food names, for example. But the use of semantic expectation also depends on the recognition of lexical items, as they are the smallest units to have meaning.

By elimination then, the only promising route to exploiting higher level redundancy for error-correction seems to be through the identification of the lexical items in the signal. Fortunately, there is considerable redundancy in the lexicon, and it can itself be used to upgrade a signal.

Redundancy in the lexicon lies in the non-existence of most of the words[2] which are possible within the language. While most syllables, if they are permitted in a

[2] Hereforward, I will use the term WORD as a substitute for 'lexical item'.

language, do occur in some word or other, many do not occur as independent words, appearing only in multisyllabic words. In Table 1 (I have used a modified English spelling), note that even with highly-favored morpheme structures, quite a few possible words are missing. With less-favored types at right, most possibilities are missing.

TABLE 1

Sample of Gaps in English Lexicon

favorite types				more complex types		
Cit	Cad	Cine	Cup	Crand	Caddle	Cister
pit	pad	pine	pup	*prand	paddle	*pister
bit	bad	*bine	*bup	brand	*baddle	*bister
fit	fad	fine	*fup	*frand	*faddle	*fister
*vit	*vad	vine	*vup	**	*vaddle	*vister
mit	mad	mine	*mup	**	*maddle	mister
tit	*tad	tine	*tup	*trand	*taddle	*tister
*dit	dad	dine	*dup	*drand	*daddle	*dister
*θit	*θad	*θine	*θup	*θrand	*θaddle	*θister
*ðit	*ðad	ðine	*ðup	**	*ðaddle	*ðister
sit	sad	sine	sup	**	saddle	sister
*zit	*zad	*zine	*zup	**	*zaddle	*zister
nit	*nad	nine	*nup	**	*naddle	*nister
chit	?chad	*chine	*chup	**	*chaddle	*chister
*jit	*jad	*jine	*jup	**	*jaddle	*jister
shit	*shad	shine	*shup	*shrand	*shaddle	*shister
kit	kad	*kine	kup	*krand	*kaddle	*kister
*git	?gad	*gine	*gup	grand	*gaddle	*gister
hit	had	*hine	?hup	**	*haddle	*hister

* accidental gap
** principled gap

Where possibilities are absent, prediction is possible. Such words are redundant because missing parts can be filled in from the knowledge of the lexicon. In this way, *cardiac* is redundant because a loss of its first or its last consonant can be restored from the knowledge that there is no word *fardiac* or *cardiap*. In the same way, an initial consonant before *-rundle* must be *t*.

There are further facts which support the contention that lexical redundancy is used in speech perception. A human being can recognize a word with the end of it missing and fill in the missing part, if not so much is missing as to make it ambiguous which word was intended. He can do the same with a medial portion missing, and even with the beginning missing, with the same limitation. More interestingly, the strength of error correction due to lexical redundancy (coupled with syntactic and semantic compatibility) is demonstrated by the "phonemic restoration" experiments by R.M. Warren and R.P. Warren (1970) at the University of Wisconsin, where the subjects

could not DETECT the loss of a phoneme or a syllable when the missing portion (the *gis* of *legislature*) is highly determinate and when its absence is covered up by a natural sound (e.g., a background cough).

To utilize lexical redundancy—the accidental gaps in the lexicon—for error correction, every form used in the language must be available. An error in the input is detected by the absence of such a form in the lexicon; selection of the closest matching word corrects the error. A device to do this must either incorporate or have access to the whole lexicon of the language.

Due to the fact that any portion of a word may be missing (or in error), every segment deriving from the input must be compared against EACH segment of EACH form in the lexicon. This total comparison also provides the possibility of segmenting the signal into words (that is, lexical items) as shown in Hofmann (1972).

Because this entails an astronomical number of comparisons, I have proposed a device called a "Vigilant Memory" to do it in real time. A small simulation has been programmed and run on the University of Ottawa computer (an IBM 360/65) described in Mes (1970). Using ordinary orthography but with errors and without spaces between words, it performed more or less as expected, except that contrary to what was apparent to us then, it performed well in spite of errors in the dictionary forms.

A vigilant memory can be built or simulated for the case of speech reduced to presently extractable parameters, and should reduce the probability of error far below anything currently proposed. Inaccurate parameter extraction can be treated simply as an additional source of noise input into the system. I suspect that the increase of noise due to errors of extraction will be nearly insignificant, compared with the external noise present in running speech.

For continuous speech, a vigilant memory, or any simulation of it, must be supplemented with a rudimentary syntax. The output of a vigilant memory necessarily contains alternates, because some words are included in others (as *cat* is in *catalogue*).[3] This syntax must choose between the various alternates in such a way as to make a consistent structure for the sentence by using syntactic redundancies. It thus accounts for the different segmentations in the more or less identical repetitions in *Have the baker recheck the bakery check*.

To summarize briefly, I have argued that there is a level of lexical perception where lexical items are recognized utilizing the extensive redundancy inherent in gaps in the lexicon. For automatic speech recognition, it provides a means to lower the error rate to an acceptable level, aided only by a superficial syntax to choose between lexical alternates.

Department of Linguistics and Modern Languages
University of Ottawa

[3] Furthermore, spurious words appear from the abutment of adjacent words, as *cat* from *picnic at*.

REFERENCES

Bruce, J.P.
1956 "Effects of Context Upon Intelligibility of Heard Speech", in *Information Theory: Third London Symposium* C. Cherry ed., (Butterworth) pp. 245-52.
Mes, L.
1970 "Simulation of Vigilant Memories", University of Ottawa.
Hofmann, T.R.
1972 "The Vigilant Memory in Segmentation and Error-Correction of Speech", *Cahiers linguistiques d'Ottawa* 2:57-67.
Warren, R.M.
1970 "Perceptual Restoration of Missing Speech Sounds", *Science* 167:392-3.
Warren, R.M. and R.P. Warren
1970 "Auditory Illusions and Confusions", *Scientific American* (December).

DISCUSSION

TRUTENAU (Legon, Ghana)
I have two comments. The first is with regard to the question of whether the kind of redundancy Dr. Hofmann talked about might be a constant in human language. I am led to believe that this may not be the case, on the following grounds: in such African tone languages as have predominantly short words, there appear to be fewer gaps in the lexical utilization of possible phonological structures. This may also be the case for older Chinese.

The second point is about the mentioned access to the total lexicon of the language. One doubts whether it is at all feasible to assume that any native speaker of any language would ever have this (otherwise my work on the Gã Dictionary Project would certainly be much easier than it is). Just to give one example: in my paper at the VIIIth Congress of the West African Linguistic Society at Abidjan 1959, I mentioned a case of 14% of the lexical material in a straigth forward word-list for school use being unfamiliar to a female secondary school leaver (native speaker).

HOFMANN
I do not know much about Gã, but in the case of Chinese, it is simply not true that there is less lexical redundancy due to 'lexical packing'. While most every syllable does have a meaning (usuable for creating new lexical items), some do not (e.g., *gèi*). Moreover, most lexical items (units which bear meaning and function syntactically) have two syllables, and there is considerable redundancy among the two-syllable words. It is conceivable that there could be a language which does not allow the use of lexical redundancy in perception. It would be very interesting to show Gã to be such a language.

As for your second point, I hasten to add that what I mean by 'lexicon' is not what is meant in synchronic linguistics (the lexicon of 'the language'). Rather, I meant the total vocabulary available to the individual in his knowledge of the language.

NOOTEBOOM (Eindhoven)
Did you imply in your paper that work on automatic speech recognition which is a technological aim in itself should proceed according to a model of human speech perception?

HOFMANN
Cut out for the lack of time was a note on my assumption: because most technology is initially an imitation of a naturally-occurring process, it seems likely that automatic speech recognition will wait on some understanding of this area.

NOOTEBOOM
Have you done any perceptual experiments? It would in particular be of interest to know whether the error rate in speech perception can be predicted in some way from the statistical properties of the lexicon?

HOFMANN
I have done no formal perceptual experiments, but because the vigilant memory can be understood as a model of lexical perception in humans, I anticipate some such experiments. The statistical properties which you suggest may be one way of testing how lexical perception proceeds.