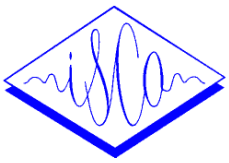
ISCA Archive
http://www.isca-speech.org/archive



2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

KNOWING ENOUGH TO ANALYZE SPOKEN LANGUAGES

Peter Ladefoged

Phonetics Laboratory, Linguistics Department, UCLA, Los Angeles, CA 90024-1543

ABSTRACT

Nobody can know everything about spoken language processing. Whatever aspect of the field we work in, we have to get help from experts in other fields. For example, for my work on the phonetic structures of dying languages, I rely on the talents of linguists, anthropologists, sound recording engineers, physiologists, psychologists and signal processing engineers to support my own phonetic knowledge. We are fortunate in that by coming to a congress such as this we can all profit and get a little help from our friends.

In the days of Leonardo da Vinci, a large library contained only a few thousand volumes. Nowadays there are tens of thousands of books that deal with speech in one way or another. Nobody, not even Leonardo, could know all that there is to know about speech research. The only way in which one can work in this field is to have good friends. We all have to rely on other people to fill in the gaps — the vast holes — in our knowledge. Any scientist today is part of a team that cannot hope to build a bridge into the future without a lot of help. This is clearly so in my case.

Much of my life is spent trying to describe how the sounds of one language differ from another. There is also some lofty goal, such as trying to develop a theory of the nature of spoken language; but that is usually in the back of my mind. In the forefront is the everyday business of describing some particular piece of spoken language. I would guess that the majority of us at this Congress could say the same thing: we spend most of our working day trying to describe some particular piece of spoken language. I concentrate on linguistic contrasts — meaningful differences — between one sound and another. One way or another, this is still true of most of us, the major exception being those who study pathologically deviant forms of speech. The rest of us are dealing with the meaningful elements of speech. In the case of some scientists, for instance communication engineers, the intent may be not to describe the meaningful elements but to process them, or to change them into other forms and transmit them. But a transformation involves a tacit description of what is to be transformed and what is not. So, in some sense we are nearly all doing the same thing, albeit in ways very different from one another.

Of course, the different aspects of spoken language that we choose to describe lead to our tasks — our daily lives — being very different from one another. But my general theme, that we have to rely on a great deal of help from many different people, is still true. My current project is the description of the sounds of dying languages. It has been estimated that about half the world's 6,000 languages will become extinct within the coming century [2]. This is a vast amount of cultural knowledge that is disappearing, and I am delighted that NSF has sponsored our UCLA research, in which we try to record for posterity the phonetic structures of some of the languages that will not be around much longer. Our first step in doing this is to identify suitable languages for investigation. This involves the cooperation of other linguists, local representatives of speakers of these languages, government officials, missionaries, and anybody who can give us advice on the viability of a particular language. We try to select languages that are phonetically interesting, for which we rely on our own expertise; but we could not do without the help of other linguists who know the particular languages. It is not practicable to go out and record the phonetic structures of a language without the assistance of someone who knows the phonology, and can give us access to a large lexicon so that we can select suitable contrastive forms.

Even more important than the help we receive from knowledgeable linguists is the help that we must have from the speakers of the language. This is somewhat outside the theme of this paper, in which I am trying to show how much we rely on other scientists in our own and neighboring fields; but it is obviously worth noting that one cannot do work on any language — dying or not — without the cooperation of its speakers. We have found that this is usually easy to obtain; most people are interested in their own language and are only too willing to share its mysteries with others. On other occasions, however, particularly when dealing with some Native American Indians, it may be more grudgingly given. For some people, language is sacred, literally god-given, and not to be casually shared with outsiders.

Having decided which language to record, our next task is to go out and do it. This, too, involves knowledge gained from other fields. Any fieldwork demands some of the skills of the anthropologist, in that one must know how to make observations within the local culture. To begin with, we must know how to choose suitable speakers who are truly representative of the population. One of the problems that I face is that local leaders expect me to record the old men and women who know the wisdom of the tribe. But these people often have weak and quavery voices; and they may not have any teeth. So, although they may be valuable in showing how people used to speak, and reliable in their control of the syntax and vocabulary of an older generation, they are not so useful in providing formant frequencies indicative of the vowels of current speech.

We must also make sure that the people we record are speaking in a normal way; we must be able to

10.21437/ICSLP.1992-1

observe the culture without disturbing it. I regret that I often have to give up on this. The exigencies of my work are such that it is impossible to spend enough time to be able to record all the phonetic structures of a language spoken in a completely natural way. I would like, for example, to be able to record all the vowels of a language in normal conversational speech. But it is just not possible to wait for each of 10 speakers to say words containing similar consonants but each of 10 different vowels, all within the context of a friendly chat. So I have to structure my observations in some way. Any advice on this topic is welcome.

These are not problems that apply to fieldwork phoneticians alone. Virtually all of us work with recorded data provided by a sample of speakers. One of the most cited early works in acoustic phonetics [7], is flawed because the authors did not pay proper attention to choosing suitable speakers. Their data do not provide a reliable account of the vowels of what they call General American English. Their speakers consisted largely of people around Bell Telephone Laboratories, New Jersey, some of whom were not even born in the United States. They cannot be considered to be a sample which is representative of any particular population.

Nor are many more recent studies without fault. The TIMIT database, sponsored by the National Institute of Standards and Technology, made a more serious attempt to control the dialects of the speakers, both by reporting the speakers' own assessments of their dialect, and by building in so-called calibration sentences that allow investigators to observe enough features to be able to place each speaker into a dialect category regardless of the speaker's own judgment. But the TIMIT database is sometimes not so successful with regard to the second point mentioned above, namely, observing the culture without disturbing it. Many of their speakers were not very good readers. As I noted, I do not know how best to record speakers talking in a natural way. I see that there are several papers on databases at this congress, and I look forward to learning something from them. And I hope that the anthropologists and sociologists amongst us will keep us all honest in these fields.

Recording a spoken language, whether in a laboratory or the outback of Australia, also demands some instrumental skills. Nowadays a good phonetician will use a DAT (digital) recorder, so as to make recordings of the highest possible quality. Regrettably many of us regard this as a comparatively simple procedure, requiring no particular skill–which is why so many bad recordings get made. We should be consulting our engineering colleagues to ensure that we are indeed using the highest quality, noise canceling, directional microphones, and the best available recording system.

We should also consult with our colleagues in physiology. We need to know about current work in speech production for many reasons, ranging from the straightforward practical help that we can get by using instrumental aids that tell us what the speaker is doing, to a deeper understanding of the nature of the units of speech that we are trying to record. My own phonetic fieldwork now relies extensively on physiological data. I cannot go out into the field with the more elaborate techniques for studying speech production described in some of the papers at this congress. There is no portable x-ray microbeam we can take into the Kalahari desert; and even techniques for studying articulatory movement such as electromyography and magnetometry are difficult to use in circumstances that require lightweight apparatus that can be dropped and bounced around in transit, and then operated without an electrical supply from a public utility. One of the new methods of studying tongue movements, electropalatography, has been successfully used in the field by Butcher (personal communication), but this technique has its limitations in that it requires the production of a special artificial palate for each speaker. Each palate costs $500 and takes about 4 weeks to produce. It would usually be too expensive and too time consuming to set up, if one is trying to record half a dozen representatives of a language in a single field trip.

Nevertheless, phoneticians are missing many opportunities if they think of fieldwork as simply involving tape recording. One can learn a lot about different places of articulation from static palatography, a nineteenth century technique in which one of the articulators is coated with a marking medium, a word is said, and then one can observe where the articulators have made contact. A video camera, which is our current technique for recording palatographic observations, will also provide useful data on labial articulations, as has been shown by my colleague Ian Maddieson [5]. Aerodynamic data is another staple of contemporary instrumental phonetic fieldwork. Butcher (personal communication) has produced some interesting studies of Australian aboriginal languages. Our recent aerodynamic studies of Sandawe, an East African click language, are described in Maddieson, Ladefoged and Sands (in press), and of Angami, a Tibeto-Burman language spoken in India, in Bhaskararao and Ladefoged (forthcoming). In all these and many other cases, the ability to record and analyze physiological data has made a valuable contribution to the description of the sounds of the language.

I suspect that, whatever our corner of the field, similar considerations apply. Whatever kind of spoken language data one is examining, one's knowledge of why it is as it is would probably be improved if physiological data were also available. Everyone who is concerned with spoken language processing should at least be aware of what our colleagues in speech production are doing. That is why we have congresses like this, so that we can all get a little help from our friends.

Nor should any of us neglect the work of our psychologist colleagues who study the perception of speech. Linguistic phonetic fieldwork has not usually involved techniques of this kind. But there has been some notable work, such as that of Traill [8], who took audiometry equipment out into the field, tested the hearing of a group of Bushmen, and then reported the results of listening tests involving clicks synthesized by different rules on a Klatt synthesizer. There is probably not much demand for speech synthesis by rule in a hunter-gatherer economy, and certainly little need for

automatic speech recognition in a society that does not use ATMs, or, for that matter, banks. But when the time comes, Traill's work will be there to help; and, more importantly, it has already provided us with useful information on auditory distinctions among clicks. We now know a little more about the perceptual phenomena that occur in the world's languages.

Other, less involved, psycho-acoustic experiments can be readily done in the field. We have a very portable system that permits subjects to find a match to a particular stimulus such as a vowel in their own language, using the protocol described by Johnson, Wright and Flemming [1]. The system can be run on any current Macintosh computer, including a power book, the only additional equipment needed being a pair of headphones. The subject's task is to use the mouse to find the best match out of 330 high quality synthesized vowels, each of which can be reproduced by clicking on one of 330 buttons arranged in a 15 x 22 matrix, corresponding to F1 and F2 values. This system has a much wider applicability than its use in our linguistic studies. It provides a way of describing a wide variety of vowels (not including nasalized vowels, rhotacized vowels, and other vowels with special considerations involved) in terms of a set of standard vowels. Subjects are usually in fairly good agreement on what constitutes the best match to a given vowel. It is possible to regard the system as an alternative to a phonetician's descriptions in terms of cardinal vowels. Using this system, phoneticians have a meaningful way of communicating with one another when they say, for example, that the mean match to the Mexican Spanish vowel [e] as in "mesa" is vowel #67, with certain formant frequencies. Perhaps those working in other aspects of spoken language processing, such as coding and ASR, might also use this same reference system, again making it possible for one part of the field to benefit from the work of another group.

Finally, while discussing perceptual psychology, I would like to appeal to this Congress and admit to a certain sense of frustration that I get when consulting some auditory psychologists concerning my practical needs. I want to know, when I describe the vowels of a language by plotting the formant frequencies, is it more appropriate to plot the formants on a mel scale, a bark scale, and ERB scale, or any other scale? None of the auditory experts seem able to agree. Some of them are distinctly unhelpful, by suggesting that I should not represent vowels in terms of formant frequencies at all; but they do not go on to say how I should represent them so that I can most usefully compare the vowels of one language with another. I suppose I will just have to go on with the ad hoc devices and intuitions I now use [3]. But if there is some general agreement on a better sytem, I would like to know about it.

Of course, the area of acoustic analysis is where most of us here overlap. There are many papers in this area that I as a linguistic phonetician have to take into account. Like most of us, I need to know about the latest analysis systems that our engineering colleagues are producing. Although we now routinely perform some analyses while we are out in the field and still have access to speakers, the bulk of the analysis of the fieldwork data is done on laboratory instrumentation when we return. We need to know the best ways of extracting descriptive parameters from our recordings. We also need some understanding of the engineering concepts involved in digital signal processing. We need to know why, for example, a 12th order LPC is appropriate when determining formants in data sampled at 10,000 Hz, and what is the best window to use when trying to measure pitch.

Sound spectrograms of one kind or another still provide a useful way of representing speech data in a visual form. Many of us from all different parts of the field are interested in what our colleague are saying about labeling spectrograms. They may be doing this for the purpose of creating large databases, but the problems they face are the same as we all have to consider when trying to interpret the facts about some particular piece of spoken language.

Moreover, databases themselves are becoming of more and more interest to many of us. We in phonetics have been using some kinds of databases for many years. Maddieson [4,6] showed how much we could learn by studying the segmental inventories of a carefully chosen sample of languages. More recently phoneticians have started making analyses of speech databases; two of my own colleagues, Patricia Keating and Dani Byrd, are among the many reporting at this Congress.

I seriously believe that the use of large databases may completely transform phonetics, phonology, and perhaps even the whole of linguistics over the next few years. Since the advent of Noam Chomsky, the emphasis in linguistics has been on describing a speaker's competence — the mental structure of language — rather than the actual performance. Linguists, even phoneticians, have tended to ask "Could one say so and so," rather than to observe what percentage of people actually say the equivalent of "so and so." I am not meaning to imply that none of the recent advances in linguistics have been data driven. Many linguists are brilliant observers of what people do, and they have elaborated their theories so as to account for their observations. Nor am I meaning to imply that the mentalist approach to linguistic description is incorrect. Of course language can be usefully described as a set of rules in a speaker's mind. But that is not the only valid, nor even the only interesting, description of the social phenomena we call spoken language. Now that we are beginning to build up large databases, we can take a different approach. It would be foolish to go back to the American linguistics of the early 1950's, when descriptions of a language were supposed to encompass all and only that which was in a corpus. But it is equally foolish to continue sitting in an armchair and pontificating about the spoken language inside some imaginary speaker's head. I know that in saying these things at this Congress I am in some senses preaching to the converted. We are a fairly data driven lot. But let us make the breadth and excellence of our descriptions of spoken language so great that none of our more theory-bound colleagues will be able to disregard us.

I started off this paper by asking what do we need to know in order to work in spoken language processing, and I hope fairly rapidly demonstrated that there was no way any of us could know it all. Let me end on a more upbeat note. Suppose we ask instead, what does one need to know in order to do *good* work in spoken language processing. I have asked a number of well known people who are coming to this meeting whether there were any out of a list of topics connected with spoken language processing that they knew nothing about. Virtually everyone who answered admitted that they knew almost nothing about at least one of the listed topics of this congress. You would probably be surprised at the confessed ignorance of some of the major figures in the field. So obviously it is not necessary to know all about spoken language. You can do good work in the field knowing only your own little corner. But it is also true that the leading figures in the field do have at least some knowledge of many different parts of it. So here at this Congress is our opportunity to fill some of the gaps. Of course we will have the usual problems. In my case I hope I will be able to hear about new things such as magnetometer sensing systems, and new descriptions of tongue movements. But it seems that I will have to listen with only one ear to each; and my two ears will have to be in separate rooms. As there are many other cases like this, I expect I will often be listening with only half an ear. There is a wealth of material awaiting us.

References

[1] J. Johnson, R. Wright and E. Flemming. "Using the method of adjustment to study vowel spaces." *Journal of the Acoustical Society of America,* vol. 91, p. 2387, 1992.

[2] K. Hale, K. Michael , L. Watahomigie, A. Y. Yamamoto, C. Craig, L. M. Jeanne, and N. C. England. "Endangered languages." *Language*, pp. 68. 1-42, 1992.

[3] P. Ladefoged. *A Course in Phonetics* (Third edition ed.). New York: Harcourt, Brace, Jovanovich, 1992.

[4] I. Maddieson. *Patterns of Sounds*. Cambridge: Cambridge University Press 1984.

[5] I. Maddieson. "Revision of the IPA: Linguo-labials as a test case." *Journal of the International Phonetic Association,* vol. 17, pp. 26-30, 1987.

[6] I. Maddieson and K. Precoda. "Updating UPSID." *UCLA Working Papers in Phonetics*, vol. 74, pp. 104-111, 1990.

[7] G. E. Peterson and H. Barney. "Control methods used in a study of the vowels." *Journal of the Acoustical Society of America,* 24, 175-184,1952.

[8] A. Traill. "The perception of clicks in !Xóo." In D. Dwyer (Ed.), *Proceedings of the 23rd. African Languages Conference,* 1992.