attributes of the ear that in a straightforward manner differentiate signals from the front and rear. Many animals have the ability to localize a sound source by wiggling their ears, humans can instead move themselves or the head to explore the sound source direction (Wightman & Kistler, 1999). As mentioned earlier the outer ear is however of great importance for locating a sound source, the shape of the pinnae does enhance sound from the front in certain ways but it takes practice to make use of such cues. In the same way the shape of the pinnae can be of importance for locating sound sources in the medial plane (Gardner & Gardner, 1973; Musicant & Butler, 1984). Subtle movements of the head, experience of sound reflections in different acoustic settings and learning how to use pinnae related cues are some solutions to the front-back-up-down ambiguity that could be adopted also by the robot. We should not forget though, that humans always use multiple sources of information for on-line problem solving and this is most probably the case also when locating sound sources. When we hear a sound there is usually an event or an object that caused that sound, a sound source that we easily could spot with our eyes. So the next question we need to ask is how important vision is in localizing sound sources or in the process of learning how to trace sound sources with our ears and how vision can be used in the implementation of directional hearing of the robot.

5 Concluding remarks

Directional hearing is only one of the many aspects of human information processing that we have to consider when mimicking human behaviour in an embodied robot system. In this paper we have discussed how the head has an impact on the intensity of signals at different frequencies and how this principle can be used also for sound source localization in robotics. The signal responses of two types of microphones were tested regarding HRTF at different azimuths as a first step of implementing directional hearing in a humanoid robot. The next steps are designing outer ears and formalizing the processes of directional hearing for implementation and on-line evaluations (Hörnstein et al., 2006).

References

Beira, R., M. Lopes, C. Miguel, J. Santos-Victor, A. Bernardino, G. Metta et al., in press. Design of the robot-cub (icub) head. *IEEE ICRA.*

Feddersen, W.E., T.T. Sandel, D.C. Teas & L.A. Jeffress, 1957. Localization of High-Frequency Tones. *Journal of the Acoustical Society of America 29*, 988-991.

Gardner, M.B. & R.S. Gardner, 1973. Problem of localization in the median plane: effect of pinnae cavity occlusion. *Journal of the Acoustical Society of America 53*, 400-408.

Gelfand, S., 1998. *An introduction to psychological and physiological acoustics.* New York: Marcel Dekker, Inc.

Hörnstein, J., M. Lopes & J. Santos-Victor, 2006. Sound localization for humanoid robots – building audio-motor maps based on the HRTF. *CONTACT project report.*

Musicant, A.D. & R.A. Butler, 1984. The influence of pinnae-based spectral cues on sound localization. *Journal of the Acoustical Society of America 75*, 1195-1200.

Pickles, J.O., 1988. *An Introduction to the Physiology of Hearing.* (Second ed.) London: Academic Press.

Shaw, E.A.G., 1974. Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *Journal of the Acoustical Society of America 56.*

Shaw, E.A.G. & M.M. Vaillancourt, 1985. Transformation of sound-pressure level from the free field to the eardrum presented in numerical form. *Journal of the Acoustical Society of America 78*, 1120-1123.

Wightman, F.L. & D.J. Kistler, 1999. Resolution of front—back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustical Society of America 105*, 2841-2853.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 49–52

Microphones and Measurements

Gert Foget Hansen[1] and Nicolai Pharao[2]
[1]Department of Dialectology, University of Copenhagen
gertfh@hum.ku.dk
[2]Centre for Language Change in Real Time, University of Copenhagen
nicolajp@hum.ku.dk

Abstract

This paper presents the current status of an ongoing investigation of differences in formant estimates of vowels that may come about solely due to the circumstances of the recording of the speech material. The impact of the interplay between type and placement of microphone and room acoustics are to be examined for adult males and females across a number of vowel qualities. Furthermore, two estimation methods will be compared (LPC vs. manual). We present the pilot experiment that initiated the project along with a brief discussion of some relevant articles. The pilot experiment as well as the available results from other related experiments seem to indicate that different recording circumstances could induce apparent formant differences of a magnitude comparable to differences reported in some investigations of sound change.

1 Introduction

1.1 Purpose

The study reported here arose from a request to evaluate different types of recording equipment for the LANCHART Project, a longitudinal study of language change with Danish as an example. One aim of the assignment was to ensure that the LANCHART corpus would be suitable for certain acoustic phonetic investigations.

1.2 Pilot experiments – choosing suitable microphones for on-location recordings

Head mounted microphones were compared to the performance of a lapel-worn microphone and a full-size directional microphone placed in a microphone stand in front of the speaker (hereafter referred to as a studio microphone). The following four factors were considered in the evaluation of the suitability of the recordings provided by the microphones: 1) ease of transcription and 2) segmentation of the recordings as well as estimation of 3) fundamental frequency and 4) formants using LPC analysis.

Simultaneous recordings of one speaker using all three types of microphones formed the basis for a pilot investigation. Primarily, the results indicated that the lapel-worn microphone was clearly inferior to the other two types with regard to the first 3 criteria, since it is more prone to pick up background noise. The head mounted and studio microphones also showed some differences with regard to these 3 criteria; in particular the recordings made with the head mounted microphone provided clearer spectrograms. Furthermore, apparent differences emerged in the LPC analysis of the vowels in the three recordings.

To explore this further we recorded 6 different pairs of microphone and distance combinations using a two channel hard disk recorder. Microphones compared were Sennheiser ME64, Sennheiser MKE2 lavallier and VT600 headset microphone, positioned either as indicated by type, or as typical for ME64 (i.e. at a distance of about 30 cm).

One speaker producing various sustained vowels was recorded, and subsequently we measured the formant values at the same 3 randomly chosen points in each vowel in the two channels and compared the values. Of course we expected some random variation, but our naïve intuition was that if a formant value would for some reason bounce upwards in one recording it should also do so in a synchronous recording made with a different microphone set up. We were wrong. In fact, vowels of all heights and tongue positions seemed to exhibit quite dramatic differences, but the differences appeared to be more prominent for some vowels. Some of these differences are likely to be the result of mistracings of formant values in one or both channels, but some of the large differences were found for high front non-rounded vowels like [i] and [e] where first and second formants are not often confused. Furthermore, when we compared average values of the three points in each vowel, 37 out of 252 values differed between 5 and 10%, while 31 differed more than 10%. Closer inspection of the two channels revealed that a substantial number of the differences could not simply be attributed to spurious values, but was indeed a result of the LPC algorithm producing consistently different results, although the average differences were of a smaller magnitude. Since all other factors were held constant in these pairwise comparisons the apparent differences could only be an effect of the type and placement of the microphone. The question remains which recording to trust.

2 Previous investigations

In previous investigations of the usability of portable recording equipment for phonetic investigations and the reliability of LPC-based formant measurements made on such recordings the main focus seems to have been on the recording devices, notably the consequences of using digital recorders that employ some sort of psychoacoustic encoding such as MiniDisc and mp3 recorders, rather than on the role of the microphone used. Below is a brief summary of the articles we have found which deal with the influence the microphone exerts.

Though the goal for van Son (2002) is also to investigate the consequences of using audio-compression, interestingly, van Son uses the difference in estimation values that results from switching from one particular microphone to another as a yardstick against which the errors introduced by the compression algorithms are compared. Comparing a Sennheiser MKH105 condenser microphone against a Shure SM10A dynamic headset microphone he finds differences between the two recordings larger than 9 semitones (considered "jumps") in slightly less than 4% of the estimates of F1 and about 2% for F2. When these jumps are excluded the remaining measurements show an increased RMS error of about 1.2 to 1.7 semitones as a result of switching microphones. Unfortunately it is not possible to see the values for the individual vowel qualities.

Plichta (2004) also examines formant estimates of vowels from three simultaneous recordings. Comparing three combinations of microphone and recording equipment (thereby not separating characteristics of the microphones and the recording equipment), he shows significant differences in F1 values and bandwidths between all three recording conditions. His material is limited to non-high non-rounded front vowels, plus the diphthong [ai].

Thus there is evidence that recordings made with different microphones (and/or recording equipment differing in other respects) can lead to significantly different formant estimates.

Apart from these investigations there are two studies of the spectral consequences of differences in microphone placement by the acoustician Eddy Bøgh Brixen (1996; 1998) which are of particular relevance to our investigation. He provides evidence that the placement of the microphone relative to the speaker in and of itself can lead to substantial differences in the recorded power spectrum, notably when microphones are placed very close to (or on) the body or head of the speaker as is the case with lavallier and headband microphones.

3 The experiment

3.1 Research questions

As we have seen recordings will be affected by a number of factors which interact in complex ways making for a source of error of unknown impact on formant estimates. Now the interesting question is: how big is the problem? Is it large enough to have practical consequences for the use of LPC-based formant estimation as an analysis tool? This overall question led us to these research questions:

a) How accurate can LPC-based formant estimates be expected to be?
b) How much does the microphone and its placement contribute to the inaccuracy?
c) How much does the room contribute to the inaccuracy?
d) Is this only a concern for LPC-based formant estimates, or are estimates made by hand also affected?

3.2 Experimental design

As an attempt to answer these questions, a more comprehensive experiment was designed. It seems to us that what we need is some sort of neutral reference recording and knowledge about the consequences for formant estimation as we deviate from this ideal. Thus we planned to compare formant estimates of recordings made in four locations with very different acoustic characteristics using four different microphones simultaneously. In total the recorded material covers: 4 microphones (see table below), 4 locations: Anechoic chamber, recording studio, two private rooms, 2 male and 2 female adult speakers.

The subjects read short sentences producing 6-18 renditions of 8 vowel qualities at each location. In addition 6 repetitions of sustained vowels with f0-sweeps of 6 vowel qualities have been recorded in the recording studio and in the anechoic chamber by four speakers. These were meant to facilitate a more accurate manual estimation of the formant values. All material was recorded using synchronized Sound Device SD722 hard disk recorders at 24 Bit/48KHz.

Table 1. Microphones compared and their position relative to the subjects

Microphone	Position relative to speaker	Directional sensitivity
Brüel & Kjær 4179	80 cm directly in front of speaker's mouth	omnidirectional
Sennheiser MKH40	40 cm at a 45 degree angle	cardioid
DPA 4066	2 cm from corner of mouth, head worn	omnidirectional
VT 700	2 cm from corner of mouth, head worn	omnidirectional

We would suggest using the B&K 4179 with a (certified) flat frequency response in the anechoic chamber at a distance of 80 cm as the reference. The distance is perhaps somewhat arbitrary, but it appears from Brixen (1998) that the effect of changing the distance decreases rapidly as the distance increases. On-axis, the spectrum at 80 cm deviates less than +/- 2dB from the spectrum at 1 m.

4 Current status and preliminary results

All planned recordings have been made, and the analysis phase has commenced. We have started with the sustained vowels as they should be the simplest to analyse (since there are no transitions to be aware of) and as they are also the most suitable for manual inspection. Two PRAAT scripts have been produced for the analysis. One is a formant analysis tool that enables simultaneous analysis of the four recordings to ensure that measurements are made at

points that – as far as possible – provide trustworthy formant values for all recordings. The other is a script which by tracing the intensity variation in each partial as the f0 changes, can be used to determine when a given partial crosses a formant. By measuring f0 at this point and counting the number of the partial we are able to estimate the formant frequency. We assume that this approach will be more accurate than judging the formant frequencies by visual inspection alone.

It is obvious that the "f0-sweep" approach we use to determine formant values manually is not without flaws as we are relying heavily on a number of assumptions: First we expect our speakers to be able to produce the same vowel quality independent of pitch. As vowel quality and pitch are known to be interrelated in real speech it may both be difficult for our speakers to live up to this expectation, and difficult for us to verify auditorily whether they do. Even if the speakers may succeed in 'freezing' the oral cavities during the sweep, differences may arise due to movement of the larynx as the pitch is changed, as well as due to changes in voice quality associated with the pitch. Notably the voice often seemed to get more breathy and hypofunctional towards the lower end of the pitch range. The method of determining the time of the maximum energy for the partial may also be affected by overall changes in intensity that have nothing to do with the interaction between the partial and the formant. This would mostly affect estimates of F1 as the transition of partials through higher formants happens much faster, and since there are often more partials crossing through the formant thus giving more estimates. Finally the accuracy of course depends on the accuracy of the f0 tracing, and more so the higher the partial. Despite the potential shortcomings of the method it does seem to provide reliable results, and is particularly helpful in determining the formant frequencies in the lower region of the spectrum.

Our ongoing analyses of the data have so far only confirmed the usefulness of carrying out the larger investigation. We hope to be able to ensure that our colleagues at the LANCHART Project need not end up reporting as sound changes what might merely be the results of microphone changes...

References

Brixen, E.B., 1996. Spectral Degradation of Speech Captured by Miniature Microphones Caused by the Microphone Placing on Persons' Head and Chest. *Proceedings AES 100th Convention.*

Brixen, E.B., 1998. Near Field Registration of the Human Voice: Spectral Changes due to Positions. *Proceedings AES 104th Convention.*

Plichta, B., 2004. Data acquisition problems. In B. Plichta, *Signal acquisition and acoustic analysis of speech.* Available at: http://bartus.org/akustyk/signal_aquisition.pdf.

van Son, R.J.J., 2002. *Can standard analysis tools be used on decompressed speech?* Available at: http://www.fon.hum.uva.nl/Service/IFAcorpus/SLcorpus/AdditionalDocuments/CoCOSDA2002.pdf.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 53–56

Prosodic Cues for Interaction Control in Spoken Dialogue Systems

Mattias Heldner and Jens Edlund
Department of Speech, Music and Hearing, KTH, Stockholm
{mattias|edlund}@speech.kth.se

Abstract

This paper discusses the feasibility of using prosodic features for interaction control in spoken dialogue systems, and points to experimental evidence that automatically extracted prosodic features can be used to improve the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor, as well as to shorten system response times.

1 Introduction

All spoken dialogue systems, no matter what flavour they come in, need some kind of interaction control capabilities in order to identify places where it is legitimate to begin to talk to a human interlocutor, as well as to avoid interrupting the user. Most current systems rely *exclusively* on silence duration thresholds for making such interaction control decisions, with thresholds typically ranging from 500 to 2000 ms (Ferrer, Shriberg & Stolcke, 2002; Shriberg & Stolcke, 2004). Such an approach has several drawbacks, both from the point of view of the user and that of the system. Users generally have to wait longer for responses than in human-human interactions; at the same time they run the risk of being interrupted by the system, since people frequently pause *mid-speech*, for example when hesitating or before semantically heavy words (Edlund & Heldner, 2005; Shriberg & Stolcke, 2004); and using silent pauses as the sole information for segmentation of user input is likely to impair the system's speech understanding, as unfinished or badly segmented utterances often are more difficult to interpret (Bell, Boye & Gustafson, 2001).

Humans are very good at discriminating the places where their conversational partners have finished talking from those where they have not – accidental interruptions are rare in conversations. Apparently, we use a variety of information to do so, including numerous prosodic and gestural features, as well as higher levels of understanding, for example related to (in)completeness on a structural level (e.g. Duncan, 1972; Ford & Thompson, 1996; Local, Kelly & Wells, 1986).

In light of this, the interaction control capabilities of spoken dialogue systems would likely benefit from access to more of this variety of information – more than just the duration of silent pauses. Ultimately, spoken dialogue systems should of course be able to combine all relevant and available sources of information for making interaction control decisions. Attempts have been made at using semantic information (Bell, Boye & Gustafson, 2001; Skantze & Edlund, 2004), prosodic information and in particular intonation patterns (Edlund & Heldner, 2005; Ferrer, Shriberg & Stolcke, 2002; Thórisson, 2002), and visual information (Thórisson, 2002) to deal with (among other things) the problems that occur as a result of interaction control decisions based on silence only.