CONSTRAINTS ON THEORIES OF HUMAN VS. MACHINE RECOGNITION OF SPEECH

Roger K. Moore
20/20 Speech Ltd, Malvern, UK
and **Anne Cutler**
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ABSTRACT

The central issues in the study of speech recognition by human listeners (HSR) and of automatic speech recognition (ASR) are clearly comparable; nevertheless the research communities that concern themselves with ASR and HSR are largely distinct. This paper compares the research objectives of the two fields, and attempts to draw informative lessons from one to the other.

1. INTRODUCTION

Speech recognition researchers aim to understand the process of extracting linguistic information from an acoustic signal. The central issues in the study of speech recognition by human listeners (HSR) and of automatic speech recognition (ASR) are, in this respect, clearly comparable; nevertheless, the research communities that concern themselves with ASR and HSR are largely distinct. In part this follows from the way the two research tasks are envisaged. In ASR, the objective has been to define a complete end-to-end process for moving from low-level to high-level descriptions of a speech signal. ASR models are typically presented in the form of an entire working system that takes acoustic input and produces a lexical or semantic-level description as output. In contrast, HSR research has been subdivided, and its separate components parcelled out to different research domains. As a consequence, HSR researchers' theories are models of specific aspects of the process, with minimal attention to the feasibility of interfacing to models of other stages in the recognition process.

We begin this contribution by comparing the research objectives of the two fields, after which we consider points of potential contact between them.

2. ASR RESEARCH OBJECTIVES

The past twenty years have witnessed a substantial growth in the capabilities of automatic speech recognition, first in the research laboratory and subsequently in the commercial marketplace. The technology has reached a point where large-vocabulary speaker-independent continuous speech recognition (LVCSR) is now available for only a few tens of dollars in any high-street computer store, and where small-vocabulary voice command-and-control is becoming a familiar feature for users of telephone-based interactive voice response (IVR) systems. These developments are the result on the one hand of the introduction of hidden Markov modelling in the 1980s, and on the other hand simply of the relentless increase in desktop computing power which has taken place in the same two decades.

The fundamental principle underlying contemporary automatic speech recognition has been definition of a methodology for moving from the lowest level physical description to the highest-level abstract description of a speech signal. For this it has been necessary to focus on:

- the a-priori information (knowledge) that is available in the form of descriptive data about speech patterns, linguistic structures and the relationships between different levels of description, together with corpora of recorded speech/language material and their annotations,
- the representation (encoding) of speech knowledge and information derived from actual speech data, and
- the computation (algorithms) that must be executed in order to achieve the required transformations.

The key issue in ASR is the nature of the encoding. What is required is a mathematical and scientific formalism for encoding a-priori information in a computationally useful form: a formalism which can exploit regularities (patterns) in the data, can generalise from seen data to unseen data (in order to perform recognition), and in response to some efficiency criterion uses the minimum information to achieve its goals.

This leads naturally to an approach that is founded on information theory and on speech pattern modelling. Information about speech and speech patterns is encoded in a suitable model, and appropriate algorithms compute the most likely state of the model for a specified input.

Current ASR is thus based on a sound mathematical formalism - hidden Markov modelling - for integrating prior acoustic, phonetic and linguistic knowledge (including context dependency) into a single integrated sequential data structure. There exists a powerful mathematical optimisation algorithm - the Baum-Welch algorithm - for estimating the parameters of the model(s) given example data. There exists a mathematical definition of recognition (what should be computed) as the most likely interpretation of the observations given the overall model - derived from Bayes theorem. And there exists an efficient (and optimal) algorithm - the Viterbi algorithm - for recognition that can be implemented as an integrated search over all constraints.

These principles lead to a set of behaviours - some emergent (that is, not programmed explicitly) - that are not only important in their own right, but which may also have interesting parallels with the traits exhibited by human listeners. These issues are addressed in section 4.

3. HSR RESEARCH OBJECTIVES

The challenge to human speech recognition research is not designing a system that will work, but fathoming the design of an extant functioning system. The major obstacle to the HSR researcher's task is the impossibility of direct observation. HSR researchers cannot take the system apart, neither can they switch off components one by one to identify their separate contributions, nor can they adjust the system and compare the performance of versions which differ only in a single respect.

The devising of methods for indirect observation is therefore one of the central arts of HSR research. Task analysis is of major importance: exactly what aspects of the recognition process are involved in performing a given task, and how by comparison across tasks can each aspect be observed separately. HSR is studied in experimental laboratories, and it is vital to maintain control over the experimental conditions, and to avoid confounding of potential contributory factors.

In good part because of the considerable expertise necessary to conduct highly controlled experimental studies, HSR research has become a thicket of differing specialisms. Some aspects of the task are considered so separate that they have traditionally been the domain of quite different research disciplines (housed, for example, in different university departments). Thus, for instance:

- the perceptual extraction of low-level acoustic information is studied by psychoacousticians or audiologists. Their research borders on areas of medical research (e.g. hearing impairment), and speech is merely one kind of auditory input that the human auditory system is confronted with. Psychoacoustic experiments very often involve simple judgements about sounds, e.g. whether two sounds are the same or different.
- the identification and discrimination of speech sounds has largely been the domain of phoneticians focussing on speech perception. They usually work in linguistics departments, are typically also informed about articulatory phonetics, and very often include a cross-linguistic dimension in their research. Phonetic experiments often involve categorisation - e.g. is a given input this phoneme or that phoneme?
- the understanding of words and utterances is studied by psycholinguists, based in psychology departments and usually specialising further in (a) the comprehension either of spoken or of written language, and (b) the recognition either of words or of syntactic and/or semantic and discourse structure. It is thus common to find a psycholinguist who works only on reading of isolated words, for instance. Psycholinguistic experiments often involve the measurement of reaction time to perform some judgement, e.g. to decide whether an input is a real word or not.

Thus by far the majority of research findings in HSR have come from behavioural experiments. Nature occasionally presents the HSR researcher with "experiments" involving damage to the recognition system, but these offer only very crude views of the system components. The principal value of language impairment studies is that dissociations may be found: one aspect of processing remains intact even though another aspect is impaired. This allows the researcher to hypothesise that particular components of the system are separate. Brain imaging techniques also allow separation of components to be inferred, if, for instance, different cortical areas are observed to be involved in different aspects of processing. Again, careful task analysis is necessary to interpret the results of imaging studies.

All types of HSR experiment are guided by models of aspects of the recognition process. Because of the fragmentation of the research undertaking, no complete model of the recognition process exists. Instead, the models are designed with a view to experimental test; they account for known findings on one aspect of the system (e.g. word recognition), and generate further explicit predictions that are put to empirical test. Virtually all models are computationally implemented.

Integration across HSR research domains is fairly rare. Word recognition models are often neutral as to the exact form of the input to the word processing system, for example, while phoneticians likewise do not commit themselves to claims about whether explicit phonemic identification is actually part of the HSR system.

4. INFORMATIVE LESSONS?

4.1. Implications of ASR for HSR

4.1.1. Statistics

"It isn't that statistics is a religion. It is that nothing better is known." - Fred Jelinek (1996)

It is clear from ASR research that it is vital to have a formal model of not only what is known, but also what is not known - so-called 'ignorance modelling'. This is needed both to provide a natural mechanism for generalising, and to be able to describe the overall process within a single unified mathematical framework.

Recognition schemes that are based on arbitrary distance/similarity metrics may or may not have relevant properties on a global scale, whereas statistical modelling has the huge advantage that such properties can be understood and exploited in appropriate ways. It is also the case that in the right circumstances, a distance measure can be interpreted as a probabilistic similarity measure - thus involving probabilities does not mean abandoning distances, it simply imposes a mathematical rigour which can avoid many layers of (potentially fundamentally flawed) heuristics.

4.1.2. Generalisation

The fact that ASR is based on a statistical modelling paradigm automatically facilitates generalisation to unseen data in a natural and well-defined way. In fact, a small but finite probability can be attached to even the most bizarre acoustic signal - even one unlikely to have been generated by any known physical system (sine-wave speech, for example). One consequence of this, is that, not only is ASR able to ride over a momentary loss of data, but it would appear to hallucinate with data that was well outside of its experience - that is, data on the tails of its probability distributions.

4.1.3. Hidden Behaviour

The use of hidden Markov modelling means that an automatic speech recogniser attends to the underlying (rather than surface) properties of a speech signal. If certain expected features of a speech signal were not present in the acoustic input (a deleted phoneme or a short period of masking noise, for example), then the recogniser would effectively continue to perceive the larger lexical object as if the missing data were present. The only consequence would be the localised lack of supporting evidence (manifest as reduced probability) for the overall lexical hypothesis and a greater potential for confusion with alternative overall explanations.

4.1.4. Implicit Processing Units

In the integrated approach to ASR, the interpretation of a speech signal in terms of any apparent processing units arises as a direct consequence of the implicit or explicit structures embedded in the models. Such structures - implemented as shared or tied parameters - can be viewed as patterning in the models that, ultimately, reflects the patterning (not invariance) in the signal itself. The consequence is not only the formulation of an efficient data structure capable of powerful extrapolation to unseen data, but also the provision of an automatic mechanism for focusing the recognition on those parts of the signal which contain the most relevant information.

For example, if every pattern were to be stored separately, and not combined into an integrated model, then for each measurement value within each pattern there would have to be a weighting to indicate its usefulness in making a subsequent recognition decision. However, by sharing information (parameters) between pattern classes, the integrated model is able to unify those aspects that are not useful across the different categories such that their differential effect is minimised.

4.1.5. Learning

The use of statistics and probability theory means that the process of adapting to new circumstances such as new speakers or new environments can be expressed in a mathematically cogent way. Adaptation and learning become local and global re-estimations of the model parameters - adaptation making small modifications to the values of existing parameters, and learning radically adjusting the model structures and topologies (creating new parameters). For ASR, the consequent issues are to do with convergence and stability; how could such an evolving system be guaranteed to retain old but useful structures and parameters, rather than be fatally distorted by some new piece of information?

It is also interesting to note that the balance between the acoustic modelling and the language modelling can be altered (or estimated incorrectly) and this would have an impact on the degree of attention given to the acoustic signal versus prior expectancies such as word frequency. Crudely speaking, this can be thought of as a model parameter that controls the degree of bottom-up vs. top-down processing at the lexical level, and takes different optimum values depending on the task to be performed.

4.1.6. Delayed Decisions

A consequence of viewing recognition as a search is that all decisions regarding the final outcome of the process are delayed until sufficient information is available to compute the best interpretation. This means, for example, that segmentation of the incoming signal arises as a consequence of the recognition process rather than as a precursor to it. In fact, one striking consequence of delayed decision-making is that many processes that intuitively seemed to need to precede word recognition (such as word-endpoint detection, phonetic segmentation and phonetic recognition, for example) can be performed more robustly as a by-product of the recognition process.

The process also has implications even beyond that which has been implemented in contemporary ASR systems - thus the same principle could be applied to other apparent pre-processes such as formant tracking, pitch detection, talker identification, streaming, grouping, channel compensation etc. All of these hitherto normalising procedures could be viewed as part of, and therefore as by-products of, the process of recognition - the advantage being that no one categorisation would depend on another being made correctly beforehand.

4.1.7. Reaction Time

Given that speech arrives as a left-to-right sequential signal, viewing recognition as search also means that the ambiguity in the input may be resolved sequentially, thereby facilitating a continuous process of signal input and recognition decision output. The important consequence of this is that the delay between input and output - the ‘reaction time’ - is not only variable but is also a function of the degree of ambiguity present.

Where there are many hypotheses, such as at the start of an utterance with limited contextual (language model) support, the acoustic perplexity, the ambiguity and therefore the input-output delay will be at its highest. Towards the end of a long word, or at the beginning of a highly predictable word, the ambiguity will be low and the input-output delay will be low (or even negative - the recogniser responding before the end has been uttered).

4.1.8. Layered Processing

The process of delayed decisions is an inherent consequence of viewing recognition not as a process driven either bottom-up or top-down, but as an integrated search through all of the implicit acoustic, phonetic and linguistic constraints. Although such a search may be performed as a single process, mathematically speaking it can be formally decomposed in order to make particular representations more explicit.

However, in order to do this without compromising the global mathematical definition of recognition, it is necessary to bridge the gap with a full-form lattice-style data structure that preserves all alternative interpretations at that level. If such a lattice is shortened (by computing the n-best sequences or by just choosing the best hypothesis, for example) then it will not be possible to recover from any error that is introduced as a result.

4.1.9. Speech and Non-Speech

The principles underlying ASR systems are not limited to speech - or even to a linear sequence of speech and non-speech events. In fact they represent a paradigm for interpreting and explaining entire acoustic scenes - even composite signals. In principle, the recognition process embodied in contemporary systems can compute the most likely interpretation of every part of an incoming signal. In practice, though, ASR devices use a limited set of non-speech models; so although every part of an incoming signal is interpreted, the fidelity of the explanation for unexpected sound sources may be compromised.

4.2. Implications of HSR for ASR

4.2.1. Universality vs. language-specificity

To a psychologist it seems obvious that the goal of research in all aspects of human language processing should be a cross-linguistically universal model, one which will account for the processing carried out by native speakers of any natural human language. This follows simply from the fact that humans are, at birth, universal processing devices, capable of acquiring whatever natural language they are exposed to, whether or not this is the language of their biological parents. However, there is no way to observe the universal except via the specific - any individual HSR experiment has to be carried out with speakers of a particular language, using materials in one or more particular languages. Given that languages differ on a wide range of dimensions, it is not necessarily the case that results from a particular experiment will be generalisable to recognition of other languages. Thus HSR research is constantly in danger of drawing conclusions that turn out to be language-specific and inapplicable to the universal model. Cross-linguistic comparisons are unavoidable, since only thus can universal regularities underlying language-specific manifestation be discerned.

4.2.2. Early specialisation

It is also obvious (to every language user) that command of one language does not automatically equip one to speak and understand other languages too. Learning a second language after childhood involves hard work, and the older one gets, the harder acquiring the vocabulary and structures of a second language becomes. Explaining this language-bound processing is also psycholinguists' business. Statistical processing is insufficient; studies show that late learners who have resided for decades in an adopted country still show subtle recognition deficits at many levels of processing, in comparison with speakers who learned the same language from early childhood.

4.2.3. Speech and Non-speech

The HSR system is attuned to distinguish speech from non-speech, and this may be a structural feature of the system present from birth. Discrimination between non-native phoneme classes is subject to interference from the native phonology, and hence may be very difficult, but the same acoustic distinctions may be accurately discriminated if they are perceived to be non-speech. Memory tasks are subject to disruption from irrelevant speech input while acoustically comparable non-speech input has no such disruptive effect.

4.2.4. Levels of processing

Although the fragmentation of research on HSR (by comparison with ASR) has arisen in large part for pragmatic reasons, i.e. the sheer size of the task and the accumulation of relevant knowledge over many generations of researchers, the division of the research task is not inherently counter-productive. That is, it has proven entirely possible, indeed reasonable, to study syntactic processing without at the same time paying attention to speech sound discrimination, and vice versa.

4.2.5. Modularity and feedback

One of the liveliest issues in HSR research concerns the independence of processing levels from one another. Rapid use of information from a higher level of processing (e.g. words) to resolve ambiguity at a lower level of processing (e.g. phonemes) is uncontroversial; the controversy concerns whether the higher-level information actually causes the lower-level processor to ignore the ambiguity entirely, or enables it to choose effectively between the alternative interpretations. HSR models with and without such feedback can both account for the empirical data [1] - there is no consensus.

4.2.6. Activation and competition

There is however consensus that human lexical processing is not based on search for a single correct candidate. Multiple candidate words, with full or partial support from the input, are activated; these include both aligned and non-aligned alternatives. Thus the phrase *phrase initial* may momentarily activate not only its component words but also other words which begin in the same way such as *freight* or *iniquity*, embedded words such as *fray*, *raise* or *niche*, and spurious boundary-straddlers such as *raisin*. A process of competition between the activated candidates leads to triumph for the sequence that gives the best account of the input as a whole. Under certain circumstances, the more competitors are active, the longer recognition takes. For example, s*peak* has competition from more words beginning similarly (*speed, speech, species, spiel*) than does *spoke (Spode*); this affects HSR performance.

4.2.7. Segmentation

The competition framework allows segmentation to be viewed as a by-product of word recognition; alignment is irrelevant to activation, only available support from the input counts. Nonetheless HSR is subject to additional effects that facilitate segmentation via modulation of candidate activation; e.g. acceptance of *raise* or *raisin* in *phrase initial* would leave a phrase-initial residue [f]. This could not be a word, and this improbability will cause lowering of activation for words that leave such a residue. It is important to note that this constraint is not driven from the lexicon - what matters is that [f] *could* not be a word, not that it *is* not a word. Other constraints on segmentation include probabilistic language-specific information about word boundary location (e.g. English words tend to begin with strong syllables) and syllable structure (e.g. certain phoneme sequences cannot occur within the same syllable). All these information sources exercise effects via modulation of competitor activation.

4.2.8. Flexibility

The HSR system is extraordinarily robust with respect to listening conditions, unfamiliar voices, articulatory distortion, bandpass restrictions. Human listeners adjust rapidly and easily to new accents, if necessary revising categorisation to suit (see McQueen et al., this meeting, for an experimental demonstration [2]). However this flexibility is only manifested within the native language. The HSR system is paradoxically inflexible with respect to acquisition of new sets of contrasts from another language, and moreover, this inflexibility of the listening system is grossly out of proportion to performance with written language. Even where scripts differ markedly (e.g. Japanese learners of English), written text is frequently reported as easier to understand than speech.

5. DIRECT ASR-HSR COMPARISON

ASR systems are usually assessed in terms of accuracy of performance, and are often implicitly compared with HSR systems that are assumed to perform at ceiling. HSR researchers, in contrast, investigate relative speed and accuracy of processing under varying conditions. Cutler and Robinson [3] suggested a way of adapting HSR techniques to assessment of ASR performance to enable a direct comparison; more common is simple use of the ASR standard, recognition accuracy, for HSR as well. It may be argued that the latter kind of comparison is in principle pointless, since such factors as vocabulary size are mismatched; for whatever reason, there are few such comparisons to be found in the literature. Below we describe one such study, and present a further ASR-HSR comparison in terms of the degree of exposure to speech.

5.1 Recognition Accuracy

By far the most comprehensive comparison between automatic and human speech recognition was performed by Lippmann in 1997 [4]. Lippmann compiled results from a number of well-known sources and presented comparative word error rates for a range of tasks and conditions. Figure 1 illustrates some of the key results, ranging from connected digit recognition to the transcription of spontaneous telephone speech. These results clearly indicate that ASR recognition accuracy lags about an order of magnitude behind that of HSR.

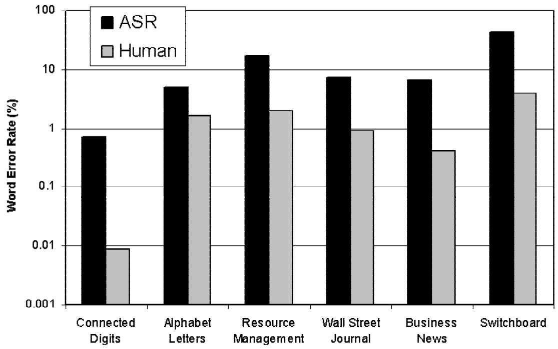

Figure 1. Comparison of human and automatic speech recognition performance (derived from Lippmann [4]).

5.2 Speech Exposure

The only comparison of the amount of data used in ASR versus HSR was presented recently by Moore [5].

5.2.1. ASR Performance as a Function of Training Data

Moore cites a paper by Lamel *et al* [6] describing 'lightly supervised acoustic model training' in which labelled training data was generated from un-annotated data using an ASR system. The application was the transcription of broadcast news material, and two conditions were studied: fully automatic annotation and annotation 'filtered' using closed-captions or transcripts. The results are presented in Table 1.

Table 1. ASR word error rates for increasing quantities of training data (taken from [6]).

Unfiltered		Filtered	
Hours	**WER**	**Hours**	**WER**
8	26.43	6	25.70
17	25.20	13	23.70
28	24.30	21	22.50
76	22.40	57	21.10
140	21.00	108	19.90

5.2.2. The Amount of Speech a Human Hears

Moore estimated the amount of speech a human hears for babies, infants and adults. Data of Van de Weijer [7] indicates that an infant receives about 20 minutes of directed speech a day. This suggested that a one year-old child would have been exposed to 130 hours of speech.

A US study by Hart & Risley [8] found that children of professional parents heard, on average, 2100 words per hour, whereas children of working-class parents heard 1200 words per hour and children on welfare about 600 words/hour. Assuming a speaking rate of 120 words per minute, an average 2/3-year old child would have been exposed to about 800 hours of speech per year. For adults, Moore assumed an average of eight hours sleep per day, and that one-quarter of the waking day is spent in conversation (i.e. two hours listening), and another couple of hours is spent listening to the radio or TV. Together these estimates suggested that an adult might be exposed to about 1500 hours of speech a year.

Moore concluded that, as a rule of thumb, a two year-old has heard 1000 hours of speech, a 10 year-old has been exposed to 10,000 hours of speech, and a 70 year-old has heard 100,000 hours of speech.

5.2.3. Comparison between ASR and HSR

Moore discovered that the data in Table 1 shows a strong linear relationship between the word error rate and the log of the amount of training material, and this meant that it was possible to extrapolate the data. The result is illustrated in Figure 2.

Moore concluded that, whilst current systems would appear to be trained on an order of magnitude less material than a two year-old infant, increasing the amount of data to that received by a ten year-old would still only reduce the word error rate of the automatic system to 12%. He also noted that the extrapolated results illustrated in Figure 2 indicate that word error rates approaching 0% would require up to 10,000,000 hours of speech training data - equivalent to 100 human lifetimes exposure to speech!

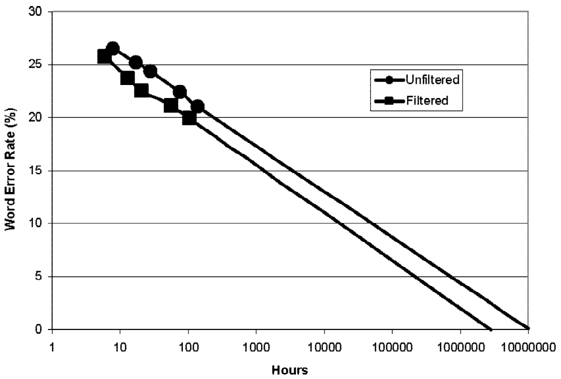

Figure 2. Extrapolated word error rates for increasing quantities of training data (taken from Moore [5]).

6. CONCLUSIONS

From the two specific comparisons between ASR and HSR, one may conclude that ASR is still behind HSR in terms of recognition accuracy, but that simply adding more and more training data is not going to provide a satisfactory strategy for ASR to catch up. What seems to be needed is a change in approach that would alter the slope of the data presented in Figure 2. In other words, true ASR progress is not only dependent on the analysis of ever more data, but on the development of more structured models which better exploit the information available in existing data. Indeed, these results suggest that the ASR research community may be squandering its data resources, thereby missing out on understanding - and thus exploiting - the underlying mechanisms which enable a human being to develop his/her astonishing listening skills in the course of only a few years.

Because the structure of the ASR and HSR research enterprises do not (as seen in sections 2 and 3 above) map easily onto one another, even ASR researchers who are interested in applying insights from HSR are often unable to do so. There is no obvious way, for instance, to use knowledge of how listeners trade cues against one another in making phoneme distinctions within an ASR system involving no phonemic-level processing. But we suggest that there are more general ways in which the two research enterprises may inform one another. The importance in HSR of the language-dependence of the system is, for instance, a potentially instructive issue for ASR researchers. If the universal HSR system is such that it becomes so tailored to the training language that it loses flexibility for non-native input, how might an ASR system benefit by being subjected to similar constraints?

7. ACKNOWLEDGEMENTS

First author's address: 20/20 Speech Ltd, Malvern Hills Science Park, Geraldine Road, Malvern, Worcs., WR14 3SZ, UK

8. REFERENCES

[1] Norris, D.G., McQueen, J.M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, **23**, 299-370.

[2] McQueen, J.M., Norris, D.G., & Cutler, A. (2001). Can lexical knowledge modulate prelexical representations over time? This volume.

[3] Cutler, A. & Robinson, T. (1992). Response time as a metric for comparison of speech recognition by humans and machines. *Proceedings ICSLP 92*, Banff.

[4] Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication,* **22**, 1-15.

[5] Moore, R. K. (2001). There's no data like more data: but when will enough be enough? *Proceedings Workshop on Innovations in Speech Processing*, UK Institute of Acoustics.

[6] Lamel, L., Gauvain, J.-L., & Adda, G. (2000). Lightly supervised acoustic model training. *Proceedings ISCA Workshop on Automatic Speech Recognition* (pp. 150-154).

[7] Weijer, J. van de (1998). *Language Input for Word Discovery.* PhD Thesis, University of Nijmegen.

[8] Hart, B. & Risley, T. R. (1995). *Meaningful Differences in the Everyday Experiences of Young American Children*, Baltimore: Paul. H. Brookes Publishing Company.