Speech is like a box of chocolates...

Helmer Strik
A^{2}RT, Dept. of Language and Speech, Univ. of Nijmegen, The Netherlands
H.Strik@let.kun.nl, http://lands.let.kun.nl/Tspublic/strik

ABSTRACT

Pronunciation variability is present in both native and foreign words. Since pronunciation variability constitutes a problem for automatic speech recognition (ASR) systems, modeling pronunciation variation for ASR has been the topic of various studies. In most studies, modeling pronunciation variation was attempted within the standard framework used in mainstream ASR systems. Given that some assumptions made within this framework are not in line with the properties of speech signals and the findings in human speech recognition, and that the improvements obtained by modeling pronunciation variation within this framework have generally been small, it might be better to look for a new paradigm in which pronunciation variation can be modeled more accurately. In this paper a novel paradigm for ASR is presented, which has many potential advantages for modeling pronunciation variation.

1. INTRODUCTION

With speech, "ya never know what your gonna get". After all, no two speech signals are identical, even if they correspond to the same utterance. This large amount of pronunciation variability is problematic for ASR [30,27,29,16,28,2]. This problem, and how it is addressed, is the topic of the current paper. Section 2 addresses pronunciation variation in native and foreign words. Special attention is paid, in section 2 and 3, to computer assisted language learning (CALL). In section 3, speech recognition by human listeners and machines is compared, and a novel approach to ASR is proposed that seems to be better suited to cope with pronunciation variation than the current approach used in ASR.

Although listeners generally have less difficulty in coping with pronunciation variation than machines do, there are cases in which pronunciation variation is problematic for human speech recognition too. Many of us will have experienced this when listening to people who speak a foreign language, like at this conference, where many participants speak English although this is not their mother tongue. If we are not familiar with the specific accent of the speaker, e.g. a Dutch, German or French accent, we may find it hard to understand what is being said. However, after listening to the same speaker for a while, we are likely to familiarize with the accent to a certain extent and we will probably be able to understand more. We may wonder why, at first, it is difficult to understand 'foreign English', and why the perception of 'foreign English' gradually improves. How are these questions and the example related to the topic of this paper? This will become clear in the remainder of this paper.

2. PRONUNCIATION VARIATION

In this section, I first discuss pronunciation variation in native language (section 2.1), and then I deal with pronunciation variation in foreign words (section 2.2).

2.1. Native language

Within the field of ASR, pronunciation variation modeling has already received considerable attention because of its potential for improving recognition performance (for an overview see [29]). Specific workshops on this topic have been organized in Rolduc in 1998 [30], Sophia-Antipolis in 2001 [16] and Estes Park in 2002 [2]; and many papers have been published (see e.g. the papers in [30,27,16,2]).

Owing to space limitations and to the fact that a lot of information on this topic can be found in the literature, here I will confine myself to mentioning that when people address an ASR-based system, their pronunciation can vary substantially, from hypo- to hyper-articulation. Many people speak in a casual way, and, especially when they pronounce frequent words, a lot of reduction can be observed. Some people, however, tend to hyperarticulate because they think it will be easier for the machine to recognize their speech. Others, who start with a 'sloppy' articulation, may switch to careful articulation within a dialog if the ASR system makes an error (see e.g. [22]).

In order to obtain satisfactory performance, the ASR system should be able to cope with this large range of pronunciation variation. To this end, techniques have been developed to adapt the acoustic models (for an overview see [32]) and the lexicon (for an overview see [28]).

2.2. Foreign words

In addition to variation in the native language, much variation can also be observed in foreign words. When people speak their mother tongue they often use foreign words or words with a foreign origin. These words can be pronounced in many different ways, which constitutes an even bigger problem for standard 'native' ASR systems. Since in many current applications foreign words (e.g. names) have to be recognized, this issue has recently received much attention (see e.g. the papers in [18]).

The problems connected to recognizing foreign words become even more acute when it comes to recognizing

ISBN 1-876346-48-5 © 2003 UAB

non-native speech, as is the case in Computer Assisted Language Learning (CALL) applications. Since it has now become common practice to integrate ASR components into CALL systems (see e.g. [8,3]), there is a growing interest in non-native speech recognition and related problems [3]. In many CALL systems the speech not only has to be recognized (in order to determine what has been said), it also has to be evaluated (in order to determine how it has been said). For good performance separate components have to be developed and optimized for these two phases: (1) recognition and (2) evaluation.

The goal of the first phase in CALL systems is to determine the content of what has been said. Recognizing non-native speech is quite difficult. The performance of standard 'native' ASR systems is too low for most CALL applications. Recognition performance can be enhanced by using ASR systems that are specially developed and optimized for non-native speech, e.g. by means of adaptation of the acoustic models and the lexicon [31,9]. In the second phase the pronunciation has to be evaluated, and pronunciation errors have to be detected. For these purposes, standard adaptation is not a good strategy. An obvious method of evaluation seems to be to compare the pronunciation of a language learner to that of a native. However, this is less straightforward than it might seem at first, amongst others for the following reasons. How to define good 'native' references, the so-called 'golden voices'? Furthermore, just storing a few references is not enough, since one cannot expect the language learners to mimic (one of) these few references. It would be better to store a very large number of 'native' reference speakers, since in this way the chances are increased that there is a reference speaker whose characteristics are close to those of the language learner. However, as far as I know, this has not yet been attempted in CALL systems. In current systems, either the students' speech is compared to a few references or a generic approach is employed in which several acoustic and temporal measures are computed [9,6]. These measures are then combined to obtain a pronunciation score or to detect pronunciation errors [9].

3. AN EPISODIC MODEL

3.1. Automatic speech recognition (ASR)

In almost all ASR systems a similar paradigm is used, which will be referred to as the mainstream ASR approach here. However, one should be aware that various assumptions made in current mainstream ASR systems are not in line with the properties of speech signals or with the findings in human speech recognition (HSR) research. A few violations are discussed below.

In ASR features are calculated locally, for fixed-width frames of 10-25 ms. The consecutive frames are assumed to be independent, which, obviously, is not the case. Furthermore, it is known that dynamics and context are important for HSR. Although information from larger time windows can be obtained from regression coefficients, the length of the time windows used in ASR is generally still smaller than the 250 - 500 ms. used by human listeners [7,14,15].

In ASR it is also assumed that speech can be described as a sequence of discrete units, usually phone(me)s. This is what I will call the symbolic approach, which has also been referred to as the 'beads-on-a-string' approach. Although this has long been one of the assumptions in (psycho)linguistics too, the idea that speech can be represented as a sequence of discrete entities proved to be untenable [12,13]. Speech is a continuous signal, continuous in the time domain and in the articulatory-acoustic domain. There are no clear boundaries in these domains. Articulatory-acoustic changes are gradual, not only between different realizations of a word, but also within a single realization. As articulators do not suddenly jump from one position to another, articulated speech sounds change gradually. Furthermore, it is not the case that at a certain point in time (the border between two speech units), all articulators suddenly change in synchrony [12,13].

The performance of state-of-the-art ASR systems is not yet good enough for many applications. The fact that almost all ASR systems are based on the same paradigm, the questionable assumptions made within this paradigm (some of which were shortly mentioned above), the insufficient performance of state-of-the-art ASR systems, and the slow progress, has led many to believe that a new paradigm for ASR is needed (see e.g. [1,28], and many papers in [25]).

3.2. Human speech recognition (HSR)

The inspiration for such a new paradigm for ASR might come from HSR. After all, many aspects that are problematic for current ASR systems are less problematic for humans, for instance pronunciation variation, noise, disfluencies, ungrammatical sentences, etc. Human beings are still much better at recognizing speech than machines [17,19]. Thus, by studying HSR one might develop ideas on how to improve ASR.

The symbolic approach used in ASR has also been used in 'traditional' (psycho -)linguistic HSR models, both in earlier and more recent publications [e.g. 26]. Furthermore, the phoneme also plays a central role in many psycholinguistic models in which a phoneme string is used as input, as e.g. in the ShortList model [20]. In the symbolic approach (both in HSR and ASR) the incoming speech signal is normalized during pre-processing, in order to strip off variation, and symbolic representations are stored in the lexicon.

However, alternative HSR models have emerged recently. These alternative models take into account that speech contains two types of information: (1) indexical information, like e.g. average F0, speech rate, intensity, and timbre, and (2) non-indexical, linguistic information. The non-indexical, linguistic information is more related to the content of the message, while the indexical information is about properties of the speaker. In the 'foreign English' example in the introduction, the indexical properties contain information about the language background of the person speaking English.

In traditional symbolic models, most of the indexical information is stripped off during pre-processing, the normalization stage. However, it is not the case that no indexical information at all is transported from the periphery to more central parts of the brain. After all, we are able to recognize familiar accents and voices (of e.g. relatives, friends, and famous people), and many can even mimic familiar accents and voices. For instance, with ‘foreign English’ one often recognizes the language background of the speaker, and some can mimic the pronunciation. So, indexical information is not discarded completely, but is perceived and encoded in memory [5].

One could then wonder whether indexical information and non-indexical information are processed independently. This turns out not to be the case. It has been observed that utterances spoken by familiar speakers are recognized better than those spoken by unfamiliar speakers [21,10]. Both indexical and non-indexical properties of speakers are stored, and familiarity with a person’s voice facilitates recognition of that person’s speech [21,10]. Many of you will have experienced this at international meetings: After you have become more familiar with someone’s ‘foreign English’, it becomes easier to understand that person.

Facilitation of recognition not only occurs with speakers that have been heard before, but also with speakers whose speech is similar. Goldinger [10] found better listening performance for speakers that had been heard before, and also for speakers that are close in a multi-dimensional space. Referring again to the ‘foreign English’ example: After having learned the ‘foreign English’ accent of one speake r, it is generally less problematic to understand the ‘foreign English’ of another person with the same language background.

3.3. An episodic model

On the basis of the findings that indexical information facilitates perception of speech by the same or similar talkers, Goldinger proposed an episodic model [10,11]. In an episodic lexicon exemplars (episodes) are stored. Whenever a new utterance has to be recognized, it is compared to all the stored tokens. If the new utterance closely resembles the stored tokens, it can be recognized relatively easily. However, if the new utterance is very different from all stored tokens (in a multi-dimensional space), the processing load will be larger, and a new token might be added to the lexicon.

Using such an episodic model for ASR has many potential advantages, some of which are mentioned here. A standard way of modeling pronunciation variation in current ASR systems is to add pronunciation variants to the lexicon. However, these variants first have to be generated, and then it has to be decided which variants should be added to the lexicon. Adding the right variants can boost performance. However, it is far from trivial to find the right variants. Furthermore, adding too many variants increases confusability, thus worsening performance. Finally, by adding variants, pronunciation variation can only be modeled at the symbolic level, while the variation is often more fine-graded. In an episodic lexicon exemplars are stored. The indexical properties can be used to determine which of the classes of speakers stored in memory is closest to the speaker that has to be recognized. Class-specific models can then be used for recognition. This will enhance performance, since it has been shown that lower word error rates can be obtained for class-specific recognition [4,24]. By using this kind of speaker (class) tracking, the need for additional adaptation of the acoustic models also diminishes. Furthermore, in an episodic memory trajectories can be stored, making it easier to use context and dynamic information during decoding.

An ASR architecture with an episodic memory might also be beneficial for CALL systems. The advantages for the recognition phase of a CALL system are similar to those described in the previous paragraph. However, there are also potential advantages for the pronunciation evaluation phase. In current CALL systems, pronunciation evaluation is carried out either by comparing the pronunciation of the student with a few stored ‘native’ references or by combining measures obtained with ‘broad’ acoustic models trained on large corpora of native speakers. The former has the disadvantage that students whose voices differ much from the stored voices will often receive low ratings. Furthermore, it will be difficult for those students to improve their pronunciation since it is difficult for them to talk like the stored references. Regarding the evaluation with ‘broad’ acoustic models, the performance is not good enough yet, while an inherent disadvantage of this evaluation procedure probably is that students speaking a mix of native voices and accents are likely to receive high ratings. This can be circumvented by using the indexical properties of the language learner.

Finally, there is one extra potential advantage. There are indications that episodic models are also applicable to visual perception. For instance, written words in an unfamiliar font are better recognized when they are presented in the context of other words written in the same new font [23]. If similar episodic models could be used for audio and visual perception, this would open up possibilities for a better integration of audio and visual perception, and, ultimately, for a comprehensive model of audiovisual perception.

ACKNOWLEDGEMENTS

I would like to thank many colleagues of my department for discussions on the topics mentioned in this paper, and their comments on previous versions of this paper.

REFERENCES

The references below are listed in alphabetical order.

[1] H. Bourlard, H. Hermansky & N. Morgan. Towards Increasing Speech Recognition Error Rates. Speech Communication, vol. 18, no. 3, 205-231, 1996.

[2] W. Byrne, E. Fosler-Lussier & D. Jurafsky (Eds.) Proceedings of the ISCA Tutorial and Research

Workshop on Pronunciation Modeling and Lexicon Adaptation (PMLA), Estes Park, Colorado, 2002.

[3] R. Carlson, C. Dunger, B. Granstrom, A. Oster (Eds.) Proceedings of the ESCA Workshop 'STiLL: Speech Technology in Language Learning', Marholmen, Sweden, May 1998.

[4] C. Chesta, P. Laface, F. Ravera. Connected Digit Recognition Using Short and Long Duration Models. Proc. of ICASSP-99, Phoenix, USA, 775-778, 1999.

[5] C.G. Clopper & D.B. Pisoni. Perception of Dialect Variation: Some Implications for Current Research and Theory in Speech Perception. In: Research on Spoken Language Processing Progress Report No. 25 (2001-2002), Indiana University, 269-290, 2002.

[6] C. Cucchiarini, H. Strik & L. Boves. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. Journal of the Acoustical Society of America, vol. 111, 2862-2873, 2002.

[7] R. Drullman, J. Festen & T. Houtgast. Effect of temporal envelope smearing on speech recognition. Journal of the Acoustical Society of America, vol. 95, 1053-1064, 1994

[8] M. Eskenazi. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. Language Learning & Technology , Vol. 2, No. 2, 62-76, Jan. 1999.

[9] H. Franco, L. Neumeyer, V. Digalakis & O. Ronen. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication* 30, 121-130, 2000.

[10]S.D. Goldinger. Words and voices: episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology; Learning Memory and Cognition. 22, 1166-1183, 1996.

[11]S.D. Goldinger. Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105, 251-279, 1998.

[12]J. Goldsmith. Autosegmental phonology. Doctoral thesis, Massachussets Institute of Technology, Cambridge, 1979. (New York: Garland Press, 1979.)

[13]J.A. Goldsmith. Autosegmental and Metrical Phonology. Oxford: Blackwell, 1990.

[14]H. Hermansky, M. Pavel. Psychophysics of Speech Engineering Systems. Proc. of the 13th ICPhS, 42-49, Stockholm, Sweden, 1995.

[15]H. Hermansky & S. Sharma. TRAPS - Classifiers of temporal patterns. Proc. ICSLP-98, 1003-1006, 1998.

[16]J-C. Juncqua, C. Wellekens (Eds.) Proc. of the ISCA Tutorial & Research Workshop (ITRW) 'Adaptation Methods For Speech Recognition', Sophia -Antipolis, France, August, 2001.

[17]D.A. van Leeuwen, L.G. van den Berg, H.J.M. Steeneken. Human benchmarks for speaker independent large vocabulary recognition performance. Proceedings Eurospeech-95, Madrid, Spain, 1461-1464, 1995.

[18]D.A. van Leeuwen, S.J. van Wijngaarden, H.J.M. Steeneken (Eds.) Proc. of the "Multilingual Interoperability in Speech Technology" (MIST) Workshop, Leusden, The Netherlands, 1999.

[19]R.P. Lippmann. Speech recognition by machines and humans. Speech Communication 22, 1-15, 1997.

[20]D. Norris. Shortlist: a connectionist model of continuous speech recognition Cognition 52, 189-234, 1994.

[21]L.C. Nygaard, M.S. Sommers, D.B. Pisoni. Speech perception as a talker-contingent process. Psychological Science, 5, 42-46, 1994.

[22]S.L. Oviatt, G. Levow, E. Moreton, M. MacEachern. Modeling global and focal hyperarticulation during human-computer error resolution. Journal of the Acoustical Society of America, 104, 1-19, 1998.

[23]T. Sanocki. Effects of font- and letter-specific experience on the perceptual processing of letters. American Journal of Psychology, 105(3), 435-458, 1992.

[24]O.E. Scharenborg, A.G.G. Bouwman, L. Boves. Connected Digit Recognition with Class Specific Word Models. COST249 Workshop on Voice Operated Telecom Services, Ghent, Belgium, 71-74, 2000.

[25]R. Smits, J. Kingston, T. Nearey, R. Zondervan (Eds.) Proceedings of the workshop 'Speech Recognition as Pattern Classification' (SPRAAC), Nijmegen, The Netherlands, 2001.

[26]K.N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. Journal of the Acoustical Society of America, vol. 111, 1872-1891, 2002.

[27]H. Strik (Ed.) Special issue of Speech Communication about 'Modeling pronunciation variation for automatic speech recognition'. Speech Communication 29, 166 pages, 1999.

[28]H. Strik. Pronunciation adaptation at the lexical level. In: J-C. Juncqua, C. Wellekens (Eds.) Proc. of the ITRW 'Adaptation Methods For Speech Recognition', 123-131, Sophia-Antipolis, France, August, 2001.

[29]H. Strik & C. Cucchiarini. Modeling pronunciation variation for ASR: a survey of the literature. Speech Communication 29 (2-4), 225-246, 1999.

[30]H. Strik, J.M. Kessens, M. Wester (eds.) Proceedings of the ESCA 'Rolduc' Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition', Rolduc, Kerkrade, May 1998.

[31]S.M. Witt, S.J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication,* 30, 95-108 2000.

[32]Phil Woodland. Speaker adaptation for continuous density HMMs: A Review. In: J-C. Juncqua, C. Wellekens (Eds.) Proc. of the ITRW 'Adaptation Methods For Speech Recognition', Sophia -Antipolis, France, August, 2001.