Benchmarking Human Performance on the Acoustic and Linguistic Subtasks of ASR Systems

László Tóth

Research Group on Artificial Intelligence,
Hungarian Academy of Sciences and University of Szeged,
Szeged, Aradi vértanúk tere 1, H-6720, Hungary
tothl@inf.u-szeged.hu

Abstract

Many believe that comparisons of machine and human speech recognition could help determine both the room for and the direction of improvement for speech recognizers. Yet, such experiments are made quite rarely or over such complex domains where instructive conclusions are hard to draw. In this paper we attempt to measure human performance on the tasks of the acoustic and language models of ASR systems separately. To simulate the task of acoustic decoding, subjects were instructed to phonetically transcribe short nonsense sentences. Here, besides the well-known superior segment classification, we also observed a good performance in word segmentation. To imitate higher-level processing, the subjects had to correct deliberately corrupted texts. Here we found that humans can achieve a word accuracy of about 80% even when almost one third of the phonemes are incorrect, and that with word boundary position information the word error rate roughly halves.
Index Terms: speech recognition, human speech perception, benchmark, performance, nonsense sentences, error correction

1. Introduction

In today's automatic speech recognition (ASR) systems the acoustic and language models are algorithmically separate and it is quite normal to train them independently. Yet, they are relatively rarely evaluated on their own - probably because from an application point of view the performance of the building blocks is of little interest. Still, we think that for the further development of ASR the behavior of the components should be analyzed and refined separately and that a comparison of their performance with that of humans would be particularly instructive. For some reason, however, such comparisons are quite rare, and in most cases fail to establish equal conditions for humans and machines, and/or do not analyze the sub-tasks in isolation [1]. In this paper we attempt to benchmark the capabilities of human subjects in solving tasks very similar to those that the acoustic and language model components of ASR systems encounter. For this purpose we designed two types of experiments. In the first one the subjects had to transcribe nonsensical continuous speech. In the second one they had to read and correct texts that were artificially spoiled with errors generated following the acoustic error pattern learned from the first experiment. Hence, the first experiment sought to evaluate the phonetic recognition performance of humans when no linguistic support is present, while the second one tried to measure the level of error correction they are capable of, relying on context and on linguistic competence. Of course, the various processing levels in humans cannot be artificially 'turned on and off' as in machines, but in spite of this difficulty we still think that the findings are quite interesting and thought-provoking.

2. Recognition of nonsense speech

The first experiment tried to mimic the task that the acoustic module of ASR systems faces: decoding the phonetic content of a continuous speech stream with no help from higher-level linguistic knowledge sources. In this section we describe the stimuli, the experiments, and then discuss the results.

2.1. Creating the stimuli

Lexicon, sentence-level syntax and semantics are handled by the language (and higher) module(s) of ASR systems, so to imitate their exclusion we had to create nonsensical test utterances. On the other hand, as the acoustic module describes the acoustic realization of the phonemes of the target language, it was also clear that the test utterances should contain only phones present in the subject's language (in this case, Hungarian). Between the lexical and phonetic levels the position of phonotactics is debatable, but we decided to count it as a part of the acoustic module for three reasons. First, context-dependent phone modelling in some sense does make use of phonotactic information. Second, in phonetic decoding ASR experiments it is usual to apply phone n-grams, which is simply a phonotactic model. Last, but not least, a fully practical issue was that a totally arbitrary series of phonetic symbols is not necessarily possible to read out loud, and this might have hindered the creation of the test data set.

To construct nonsensical words that obey the phonotactics of Hungarian, we applied a three-step procedure. First, we decided to use syllables as building blocks, and created a syllable inventory based on a large text corpus (taken from the Hungarian Electronic Library). Slightly unusually, 'syllable' here means units that go from one vowel to the next, because this way we could guarantee that only phonotactically permitted consonant clusters appeared in our words. As the second component of word construction, we created statistics of the vowel sequences occurring in the words of our training corpus. Our intention here was to model vowel harmony, which is a special feature of Hungarian and cannot be modelled at the syllable level. This way the construction of a nonsense word consisted of three steps: the generation of a vowel 'backbone', the fill-in of the gaps between the vowels by syllables, and finally a manual removal of meaningful words, should they accidentally have arisen (in fact, we tried to keep just the kind of non-words that differed in at least two phonemes from any existing word). We should mention here that the spaces between the words were

treated as special 'vowels' during modelling. This is very important because the word-leading and word-ending phone clusters obey special phonotactic rules, and this way the model was able to handle them (at least, within the range of syllables).

A final thing we had to decide on was the duration of the stimuli. Human speech perception studies usually apply very short - monosyllabic - test signals [2, 3], but we think that this kind of conditioning is unfair and not the same as that for those ASR systems that have to recognize continuous speech. First, utterances of continuous speech behave differently (e.g. show higher coarticulation). Second, monosyllabic stimuli simply do not permit the study of segmentation related phenomena (of phones or syllables or words). But, of course, using longer stimuli may give rise to errors not only due to phonetic confusions but also due to recall difficulties. The results of research on working memory and on listening span in particular suggest that people can easily repeat nonsense words of about 6-8 syllables [4]. Based on this, the sentences we created consisted of 6-10 syllables (3-4 non-words of 1-4 syllables).

2.2. Experiments

The test material contained 20 nonsense sentences, recorded on the voice of a male speaker who was instructed to read them with some arbitrary, but natural-sounding intonation. The recording also contained an introductory text to explain the task and to allow the subjects to get accustomed to the voice of the speaker and to choose a convenient loudness level. The 25 subjects were university students with no known hearing impairment, and as they were not familiar with any phonetic alphabet, they were instructed to use the Hungarian one. This caused no problem because the Hungarian orthography is almost exactly phonemic. Each sentence was repeated twice, and the subjects were told to write down their guess after each sample.

The test results were evaluated as follows. From the two responses of a subject to a given sentence the first one was kept, unless the subject committed some error that was obviously not transcriptional; for example he/she exchanged, repeated or omitted whole words (the subjects usually indicated when they failed to recall what they heard by making dots). In those cases their second reply was used, or the whole sentence was dropped when the second reply was also unusable. After this unification step the replies were evaluated in exactly the way as is usually done in ASR – that is, by matching the stimulus and the response string using dynamic programming, and then calculating a 'phoneme accuracy' score based on the number of substitution, insertion and deletion errors.

2.3. Results and Discussion

We found that the duration of the sentences rarely caused a problem, so from the two replies to a sentence we could use the first one in 92.20% of the cases, and we had to rely on the second one in only 6.80% of the cases. Both of the responses were unusable in only 1.00% of the cases.

The recognition accuracy obtained for the replies chosen this way was 83.55%. A repeated evaluation over the second responses gave a score of 87.55%. This reassured us that most of the errors are not due to memory recall, as after the second listening the subjects made only minor corrections. By way of reference, on a similar (Hungarian) recognition task our ASR system was able to attain an accuracy of 52.48% [5], and on the most studied English TIMIT corpus the best scores are around 75% [6]. These are of course not completely fair comparisons as they were measured on different data sets, but they still show the gross tendency that human performance is about 2-4 times better than today's machines. Notice, however, that the scores we got are much worse than those reported for syllabic words [3].

Besides evaluating the recognition accuracy a phone confusion matrix was also created that allows an in-depth analysis of the confusion patterns [2, 3]. Unfortunately, here we cannot discuss all the details because of lack of space. But stated in brief, we found that the most frequent source of errors was the confusion of short and long phoneme pairs and the misclassification of voicing for fricatives. This finding justified our expectations on which features were the less robust.

Although by using nonsense stimuli we tried to exclude lexical influence, it cannot be turned off completely. For example, in ambiguous cases subjects are known to be biased toward choosing lexically consistent hypotheses [7]. Examining the responses, we found 61 different meaningful words in 125 occurrences. Some of these directly corresponded to the non-words of the stimulus (eg. *vassog* → *vasfog* 'iron tooth', *pakró* → *apró* 'tiny'), but they were mostly created via incorrect segmentation. Many of these cases can be explained by the motivation to uncover a frequent real word (*kérida* → *kéri* 'asks' 17×, *metki* → *ki* 'out' 8×, *szerés* → *és* 'and' 6×) or to decompose a very unlikely consonant cluster (*ejtréled* → *ej* 'ah' 12×).

A further important aspect is that Hungarian is an agglitunative language which operates with a rich set of suffixes. As these suffixes are syllables or just consonants, a statistical model of phone sequences will inevitably describe not only phonologic, but also morphologic constraints. These suffixes are observable in our nonsense words and obviously influenced the subjects in word segmentation. For example, the nonsense sentence *Taku töhegét ötyöl* is readily analyzed by a Hungarian as consisting of a subject *taku* (having no specific suffix, so it could be a proper noun), an object *töhegét* (a noun in accusative case) and a predicative *ötyöl* (a verb in third person). In these cases morphology supposedly supports the transcription, but in other cases it might be misleading as well. For example, the word-ending [t͡ʃ] (coded by letter *cs* in Hungarian) in *albács lálolta* was decomposed by four subjects as a [t] and an [ʃ] separated by a word boundary (that is, they wrote *albát slálolta*). A probable explanation for this is that the word-candidate *albát* is morphologically more appealing than *albács*, even at the price that this way the subsequent word begins with a phonologically improbable [ʃl] pair. A further interesting example of morphologic influence was that the subjects often compensated the assimilation across word boundaries. For example, the [ʃ] in *degrás gyamar* is pronounced as [ʒ], due to regressive voicing assimilation. However, 11 subjects still decoded it as [ʃ] – apparently they identified the word boundary, and then preferred the morphologically much more probable [aːʃ] word ending to [aːʒ]. These examples show that lexical and morphologic bias was indeed present in the responses, but while the effect of the former was clearly detrimental to phonemic accuracy, on balance, the pros and cons for the latter are far from clear here.

Next, we examined whether the subjects were able to segment the utterances into words. A word boundary was judged to be correct if it separated the last and first vowels of the neighboring words; that is, the misalignment of closing or opening consonants was not judged to be an error. 32.65% of the word boundaries were visually identifiable on the spectrogram as small pauses, these were not taken into account during the evaluation. The remaining boundaries were correctly detected in 73.83% of the cases by the subjects, although these boundaries were not marked by an acoustically obvious pause. All this accords well with results which indicate that the phonetic or

phonological structure of speech contains cues about the likely location of word boundaries (for a nice survey of psycholinguistic research on word segmentation, see the dissertation by Smith [8]). For example, it is known that even infants are able to perform word segmentation at a very early stage of language acquisition by exploiting statistical regularities and prosody cues [9]. In the case of Hungarian, two recent studies have found that word segmentation could indeed be significantly supported both by prosodic cues and phoneme sequence constraints [10, 11]. Still, the current ASR technology ignores these acoustical/prelexical pieces of information and performs sentence parsing based solely on lexical fitting.

3. Context-based error correction

The goal of the second experiment was to find the phonemic error rate at which humans are still able to correct a text using higher linguistic levels. Exploiting the fact that Hungarian writing is almost exactly phonemic, we carried out this investigation in the form of a reading task of error-spoiled texts. This task can be viewed as an imitation of an ASR system's acoustic module passing a 1-best decoding task to the higher-level (lexical, syntactic, semantic, etc.) modules.

3.1. Stimulus

As test material, 27 contiguous sentences (201 words) were selected from an unpublished novel. The text was phonemically transcribed (using the Hungarian alphabet) and partitioned into 5 blocks of approximately the same size by breaking it up at sentence endings. The text was then deliberately corrupted by introducing substitution, deletion and insertion errors according to the probability distribution (confusion matrix) obtained in the first experiment. The error rate was increased block-by-block by decreasing the elements in the diagonal of the confusion matrix and distributing this probability mass proportionally among the other elements. Thus the phonemic accuracy of the blocks fell from 84% to 64% in 5% steps. 25 test sheets were generated this way, each sheet being different.

3.2. Experiments

25 new subjects were asked to figure out the text and write their guess below the sentences. It was pointed out to them that the test sentences form a sound context that is worth figuring out and exploiting. No time limit was specified and jumping about in the text was also permitted. The word boundaries were not marked in any way, but when a subject said he had finished, he received a new version of the text that contained spaces as well, and was asked to revise his guesses based on this new information. The replies were evaluated by counting the correctly found words. As insertion errors were very rare, the error rate calculated this way was practically the same as the 'word error rate' score usually applied for the evaluation of ASR systems.

3.3. Results and discussion

Figure 1 shows the word accuracy scores obtained as a function of phonemic accuracy, with and without word boundary position information. The falling tendency of the scores is just what we expected, but there are small anomalies too. For example, the readibility of the three middle blocks proved quite similar due to context, in spite of their different phonemic error rate. In fact, many subjects reported that the difficulty of the task did not grow steadily in the text, as the instructions had said. Many

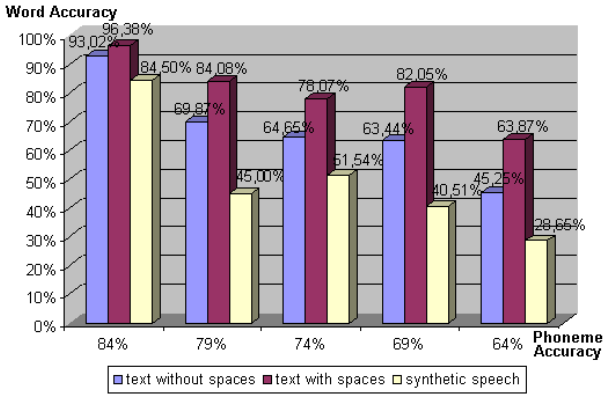

Figure 1: *Word accuracy as a function of phoneme accuracy and stimulus type.*

of them also told us that at first sight they found the task rather frustrating. Then, when they managed to figure out the outline of the story, the task became much easier. Hence, motivation turned out to be an important factor in achieving a good score, as the less motivated subjects were inclined to give up quickly.

A comparison of the first two columns in Figure 1 shows that with the knowledge of word boundary positions the error rate roughly halves. The main reason for this is obviously the reduction of entropy, and our aim with this experiment was to see how this reduction supports comprehensibility. Unfortunately, there is a further factor here that is difficult to separate or exclude. While there is still an on-going debate on exactly what information we use when we are reading, there is lots of evidence that we are able to exploit reading units (visual patterns) bigger than letters – in particular when they are highly familiar (e.g. frequent function words) [12]. The large error rate (and the lack of spaces in the first test) of our test material disturbs or perhaps sabotages this 'holistic' processing, which might explain the frustration reported by the subjects. Although analytical (i.e. letter-by-letter) reading is of course still possible, it seems that the gap between the two tasks was to some extent caused by the disturbed reading process, and not totally attributable to the different entropy levels. What supports this conjecture is the fact that the most persistent subjects managed to attain a 88-90% average word accuracy even without knowing the word boundaries (these scores went up to 93-94% with the help of spaces).

Although Hungarian writing is quite close to being phonemic, there is a possible argument against presenting the test in a written form rather than a spoken one. The test material was spoiled with typical acoustic errors based on the confusion matrix obtained in the first experiment. One might hypothesize that the auditory processing path is prepared for correcting these typical errors, while the same may not work during visual processing. That is, while the phones [k] and [g] are acoustically similar, the corresponding letters are not visually similar, and thus correcting a [k] to a [g] might seem more reasonable in a sound than in a written form. Although we instructed the subjects to try to read out the text aloud, and there is evidence that we often focus on the sound of words even in silent reading [13], we conducted a further experiment to examine this possibility. We simply took the text of the tests and fed them into a speech synthesizer (there seemed to be no difference in the synthesized signal when the text contained the spaces and when it did not, so we did not separate these cases). Ten new subjects were asked to

transcribe these sound files, allowing as much time and as many rewindings as needed. The third columns of Figure 1 show that the scores obtained in this way were significantly worse than those obtained with the written version. Although to some extent this can surely be attributed to the quality of the speech synthesizer, the results definitely do not support the assumption that the acoustic way of processing had a strong advantage due to some prior knowledge of the error patterns.

While the subjects were able to figure out more than 60% of the words even at a phonemic error rate of 36%, on the other hand it is interesting to note that they did not produce a definitive 100% even at the smallest error rate (corresponding to the error rate measured in the nonsense speech listening experiment). A possible reason for this might be that our error-generating algorithm did not model the dependence of the errors on context and position. That is, our software corrupted each phoneme independently of both its neighbors and its position in the word. In reality it is known (at least, for English) that the word onsets are more robust to assimilation and deletion than other parts of words [14], and that syllable onsets and codas also behave quite differently in this respect [15]. Humans are thought to exploit this behavior during speech comprehension [14], so it would be interesting to see what our scores would be if the error model took into account these factors as well.

As a final remark, we should mention the similarity of our experiment to Shannon's classic letter-guessing game where he attempted to assess the entropy of English [16]. However, there are at least two significant differences. First, in his experiment the subjects had to guess the next (missing) letter, while for our subjects the positions of errors were not indicated. Hence, in our case the task was *error correction* rather than *prediction*, which are clearly different. Second, Shannon dealt with printed English, while we were interested in spoken language; that is why we used phonemic transcripts. While in a discussion he mentions the possibility of repeating his experiment with phonemes instead of letters [17], we could not find any evidence that he finally did so. It is not hard to predict that the findings – for example, that "you can delete all the vowels in a passage and have no difficulty in reconstructing it" [17] – would be quite different if one had to listen to the same text via a speech synthesizer.

4. Conclusions

In this paper we attempted to measure human performance on the tasks of the acoustic and language models of ASR systems separately. In the phonetic transcription of short nonsense sentences our subjects achieved phone accuracies around 83-87%. Though the superior performance of humans over machines in phonetic classification was already known, earlier studies usually worked with syllabic input that does no allow the investigation of segmentation phenomena. In this respect we found that humans are able to hypothesize word boundaries correctly in 74% of the cases even when they are not acoustically indicated by a pause. This suggests that the suprasegmental cues currently ignored in ASR may play an important role in human speech perception and could significantly support ASR systems as well. This conjecture is reinforced by the results of the context-based text correction task where we found that the knowledge of the word boundary positions approximately halves the word error rates. In this experiment we also found that although humans can achieve a word accuracy of about 80% even at a phone error rate around 30%, they could not produce a score of 100% even at a phone error rate of only 16%. This finding and research on human lexical access [14] could indicate that our error model was overly simplified and that the distribution of errors among words and among phoneme positions within words is an important factor that should be studied in greater depth.

5. References

[1] Lippmann, R. P., "Speech recognition by machines and humans", Speech Communication, 22(1):1–15, 1997.

[2] Allen J. B., "How do human process and recognize speech?", IEEE Trans. Speech and Audio Proc., 2(4):567–577, 1994.

[3] Meyer, B., Wesker, T., Brand, T., Mertins, A., Kollmeier, B., "A Human-Machine Comparison in Speech Recognition Based on a Logatome Corpus", Proc. Workshop on Speech Recognition and Intrinsic Variation (SRIV'2006), pp. 95–100, 2006.

[4] Baddeley, A. D., Human Memory: Theory and Practice, Allyn&Bacon, 1997.

[5] Toth, L., Kocsor, A., "A segment-based interpretation of HMM/ANN hybrids", Computer Speech and Language, 21:562–578, 2007.

[6] Schwarz, P., Matějka, P., Černocký, J., "Hierarchical Structures of Neural Networks for Phoneme Recognition", Proc. ICASSP'2006, pp. 325–328, 2006.

[7] Ganong, W.F., "Phonetic categorization in auditory word perception", Journal of Experimental Psychology: Human Perception and Performance , 6:110–125, 1980.

[8] Smith, R., The role of phonetic fine detail in word segmentation, Ph.D. Thesis, University of Cambridge, 2004.

[9] Mattys, S. L., Jusczyk, P. W., "Phonotactis and Prosodic Effects on Word Segmentation in Infants", Cognitive Psychology, 38:465–494, 1999.

[10] Vicsi, K., Szaszák, Gy., "Automatic Segmentation of Continuous Speech on Word Level Based on supra-Segmental Features", Int. J. Speech Technology 8(4):363–370, 2005.

[11] Szaszák, Gy., Németh, Zs., "Word Boundary Detection Based on Phoneme Sequence Constraints", Proc. Conf. of Ph.D. Students on Computer Sciences (CSCS'2006), Vol. of Extended Abstracts, pp. 91–92, 2006.

[12] Allen, P.A., Emerson, P.L., "Holism revisited: Evidence for independent word-level and letter-level processors during word and letter processing", Journal of Experimental Psychology: Human Perception and Performance, 17:489–511, 1991.

[13] Sparrow, L., Miellet, S., "Activation of phonological codes during reading: Evidence from eye movements", Brain and Language, 81:509–516, 2002.

[14] Gow, D. W. Jr., Melvold, J., Manuel, S.,"How Word Onsets Drive Lexical Access and Segmentation: Evidence from Acoustics, Phonology and Processing", Proc. ICSLP'96, pp. 66–69, 1996.

[15] Greenberg, S., Carvey, H., Hitchcock, L., Chang, S., "The Phonetic Pattern of Spontaneous American English Discourse", Proc. Workshop on Spontaneous Speech Processing and Recognition (SSPR'2003), 2003.

[16] Shannon, C. E., "Prediction and Entropy of Printed English", Bell Systems Technical Journal, 30:50–64, 1951.

[17] Shannon, C. E., "The Redundancy of English", In: Claus Pias (ed.), Cybernetics. The Macy Conferences 1946–1953, vol. 1: Transactions, pp. 248-272, diaphanes, 2003.