# ASR "Sweet Sixteen": An Evaluation of Nuance Swedish Speech Recognizer Success Rates in 69 Commercial Applications 16 years After Its Inception and an Assessment of Inter- and Intralabeler Agreement

*Robert Eklund*[1],[2],[3]
[1]*Voice Provider Sweden, Stockholm, Sweden*
[2]*Stockholm Brain Institute/Karolinska Institute, Stockholm, Sweden*
[3]*Department of Computer Science, Linköping University, Linköping, Sweden*

## Abstract

*This paper presents an analysis of success rates of the Nuance Swedish Speech Recognizer in 69 commercial applications provided by Voice Provider Sweden. The analysis is based on 185 quality assurance reports from the period January 2007 through October 2011. An inter- and intralabeller agreement analysis is included.*

## Introduction

The starting point of the Nuance Swedish recognizer can be placed in 1995 when Telia Research (Sweden) and SRI International (Menlo Park, CA) negotiated the *Spoken Language Translator* project (Rayner et al., 2000), a speech-to-speech translation system (primarily English and Swedish), where a constituent part was the creation of a Swedish automatic speech recognizer, based on the SRI "Decipher" recognizer for US English.

Early in 1995, the author created the linguistic-phonetic training and test material for the Swedish recognizer (Eklund et al., 2000), and sixteen years later had had more than three years of "hands-on" experience of the commercial version of the Nuance recognizer at Voice Provider Sweden in a large number of services and applications. Partly the result of normal quality assurance activities, partly the result of "diachronic curiosity" on behalf of the author, the question presented itself as to how well the Swedish recognizer performs sixteen years after its first stumbling steps in 1995.

## Evaluation of HCI Systems

Modern life includes an increasing number of Human–Computer Interaction (HCI) systems, both graphical and speech-based. As a natural part of most research and development, some kind of evaluation of the created systems is carried out, and HCI systems are no exception.

A key term in the evaluation of HCI systems is "usability", but how to measure usability is not as straight-forward as might be thought. The ISO standard (ISO-9241-11) recommends that the dimensions *effectiveness*, *efficiency* and *satisfaction* should be evaluated, with summary definitions taken from Frøkjær, Hertzum & Hornbæk (2000:345) given below:

*Effectiveness*: "accuracy and completeness with which users achieve certain goals … include quality of solution and error rates"

*Efficiency*: "task completion time and learning time"

*Satisfaction*: "users' comfort with and positive attitudes towards the use of the system"

How to actually evaluate these parameters is quite another question, and in a meta-study of the literature on HCI systems evaluation Hornbaek (2006:93) found a "disarray of measures of satisfaction". For example, there is little consensus as to exactly what parameters should be evaluated and Hornbaek (2006:90) listed more than 100 words and phrases encountered in the literature, as used in questionnaire evaluations (Likert scale) of end-user satisfaction.

Moreover, several studies have shown that the three usability measures mentioned above are not strongly tied to one another, and already Walker et al. (1998:587) observed that "users' preferences are not determined by efficiency per se". Hornbaek & Law (2007) conducted a meta-analysis of 74 studies and found only a "small to medium" correlation between the three usability measures (with the weakest correlation between effectiveness and satisfaction) and Hornbaek (2006:97), in a meta-analysis of 189 previous studies, concluded that there are "notable problems in how usability measures are employed".

It would seem, then, that despite numerous evaluations of HCI systems there is still little consensus as to exactly *what* should be measured or *how* this should be measured. Not only does this make it hard to assess the usability of *particular* systems, it also makes it difficult to compare the usability of *different* systems. However, what seems clear is that different evaluation parameters do not walk in lockstep and that high user satisfaction is not automatically indicative of high efficiency or effectiveness, or vice versa.

## Evaluation at Voice Provider

Voice Provider offers a uniquely interesting material when it comes to evaluation of recognizer performance. Instead of evaluating prototype or mock-up systems in the laboratory, the Nuance recognizer is used in more than 70 different commercial services—covering a wide variety of different types of applications, including ticket reservations, technical support, travel information etc—with an associated diversity with regard to end-user characteristics.

As part of quality assurance work regular quality reports are produced where call success is evaluated. A subset of calls are extracted, listened to and labeled/classified, mainly in the *Effectiveness* dimension mentioned above. The calls are classified according to three main categories: *Successful*, *Not Successful* and *Ignored*, with short definitions given below.

*Successful*: Calls that fall within the intended functionality of the service, i.e. the end-user receives desired information (or similar) without having the repeat any one utterance more than two times in order to be understood.

*Not successful*: Calls that also fall within the intended functionality of the service, but where the end-user does not succeed in obtaining desired information (or similar). Most often this is due to failed recognition, and a prerequisite is that the sound quality, vocabulary etc all are of a quality and within a scope that the system should be expected to handle. Moreover, cases where end-users are understood, and manage to obtain the desired information (or similar) but have to repeat an utterance three (or more) times in order to be understood are also classified as not successful, according to a "rule-of-thumb" that repetitions of one and the same word/utterance severely increases user dissatisfaction with a system (Hura, 2008:207). (Note that this falls within the *Satisfaction* category above.)

*Ignored*: Calls that fall outside the intended scope and/or functionality of the service in question, or in other ways could not be expected to be handled by the system. This could be due to a huge number of reasons, including, but not limited to: prank calls; calls from sound environments that makes it hard even for a human listener to understand what is said; calls that fall outside to scope of the service; calls from inebriated end-users that fail to adhere to normal communicative behavior; calls where the end-users have misunderstood what the system is designed to cover, etc.

## Inter-Labeler Agreement Analysis: The Kappa Statistic

When several people are involved in the classification of data the obvious question is whether they use labels in the same way. Provided that a test set can be set aside that is labeled by several labelers, the most straightforward method to analyse labeler agreement would be a simple confusion matrix, but an obvious problems associated with this method is that confusion matrices do not consider agreements that occur by chance. Consequently, Cohen (1960) created a test that corrected for chance agreements, *kappa* (κ), defined as:

$$\kappa = \frac{Po - Pe}{1 - Pe} \quad \text{...where:}$$

*Po* = observed agreement between labelers
*Pe* = expected agreement between labelers
$\kappa$ = where 1 = complete agreement
and 0 = complete lack of agreement

Given the increasing need of statistical analysis of labeling agreement within the speech community, Carletta (1996) suggested that kappa testing should be employed to measure interlabeler agreement, a suggestion which was very much heeded, and a few years later Di Eugenio & Glass (2004:95) commented that "the kappa coefficient of agreement has become the de facto standard for evaluating intercoder agreement for tagging tasks".

However, kappa is both limited and prone to error. First, kappa "can compare only two encoders" (Krippendorff, 2004:413) which limits its usability when several labelers are involved. Second, depending on how symmetrical margin sums are, kappa might both underestimate and overestimate actual probability values (Feinstein & Cicchetti, 1990; Cicchetti & Feinstein, 1990).

## Data Analysis

185 quality reports covering the period 2007 to October 2011 were analyzed, as produced by eight labelers. The data set is given in *Table 1*.

*Table 1. Summary Statistics for the call data set.*

| Σ calls | 81437 (*A*) | | |
|---|---|---|---|
| *A*−*B* = | 61710 (*C*) | | Ignored (*B*) |
| Label | Successful | Not Successful | 19727 |
| Σ | 57301 | 4409 | |
| % | 92.9 | 7.1 | 24.2 |
| st. dev. | 0.06 | 0.06 | 0,15 |

Given the previously mentioned problems associated with kappa, the huge number of pairwise tests that would have been required to cover all labeler/labeler and category–labeler pairs, and the simple fact that no data existed where more than one labeler had worked on the same data, a kappa analysis was neither available or realistic.

The problem was further elevated by the fact that several of the covered services differed considerably in character and complexity and it thus cannot be taken for granted that they exhibit the same profile, and consequently elicit the same labeling behavior and/or agreement. Thus, an alternative method was chosen where test-of-proportions (*Z*) were conducted for all report/labeler pairs within each service—but not across services, given their different characteristics. This resulted in 410 *Z* tests, broken down for years and whether or not the reported *Successful* and *Ignored* rates were produced by the same or different labelers. The results of these tests were then added to different cells specifying (a) whether there was a significant difference or not; (b) whether the test compared *Successful* or *Ignored*; (c) if the test compared the *Same* labeler or *Different* labelers. These figures/proportions were then submitted to statistical analysis (χ2 tests).

A final analytical problem is that the number of quality reports that the eight labelers had produced varied substantially. Substituting labeler names with letters (A–H), the number of reports is as follows: A=90; B=69; C=13; D=5; E=2; F=2; G=2; H=2. Note that this means that 85,9% of the reports were produced by two labelers, which is likely to skew the results. This also inevitably means that intralabeler tests are more frequent than interlabeler tests.

## Results

The results are shown in *Table 2*.

*Table 2. Results for inter- and intralabeler analysis based on 185 quality reports covering 69 different services. For each service/year a Z test of proportion was carried out for all possible pairwise combinations of the two categories* Successful *and* Ignored*. This means that in cases with only one quality report no Z tests were conducted; for services/years with two quality reports (e.g. 2007 and 2009) two Z tests was conducted, one for 2007/*Successful *versus 2009/*Successful *and a second for 2007/*Ignored *versus 2009/*Ignored*; for service/years with three quality reports (e.g. 2007, 2008 and 2009) six Z tests were performed; etc. Each comparison without significant difference was added to the NSD column below, while each comparison with a significant difference was added to the SD column. The accumulated figures were then broken down for whether or not the comparisons in question belonged to the* Successful *or* Ignored *categories, and further broken down for whether or not the comparisons were made between two reports with the same labeler ("Same", i.e. intralabeler agreement) or two different labelers ("Different", i.e. interlabeler agreement). Finally,* χ2 *(Goodness-of-Fit) tests were performed to see whether there were significant differences between two categories as a function of whether or not the reports had the same labeler or different labelers.*

| Σ pairwise *Z* comparisons | 410 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Significance (95% level) | No significant difference between reports (NSD) | | | | Significant difference between reports (SD) | | | |
| Σ reports | 255 | | | | 155 | | | |
| χ2 (Goodness-of-Fit) | $p < 0.001$ | | | | | | | |
| Label category | Successful | | Ignored | | Successful | | Ignored | |
| Σ within category | 155 | | 100 | | 53 | | 102 | |
| Labeler (Same/Different) | Same | Different | Same | Different | Same | Different | Same | Different |
| Σ within category | 77 | 78 | 41 | 59 | 22 | 31 | 48 | 54 |
| χ2 (Goodness-of-Fit) | $p = 0.936$ | | $p = 0.072$ | | $p = 0.216$ | | $p = 0.552$ | |

As is shown in *Table 2*, none of the four contrasts exhibit significant differences as a function of whether the same or different labelers had made the classifications. Only in one do we observe a weak tendency towards significance in the Same/Different dimension: in the *Ignored* class in the group of no significant difference. That this category comes closest to significant differences is perhaps not surprising since the *Ignored* category is often quite arbitrary in character—e.g., judging how drunk an end-user can you be and still be expected to be understood.

## Discussion and Conclusions

Performing 410 statistical tests in order to provide the input for a batch of additional statistical tests clearly is less than ideal, but overall analysis still hints at a stable labeler agreement as far as the category *Successful* is concerned, which makes the success figure 92.9% reliable, even if obtained in a slightly "roundabout" way. Moreover, bearing in mind that this figure is not derived from laboratory tests on homogenous groups of subjects, but from a wide variety of live applications with a huge diversity along several dimensions, including end-user characteristics, acoustical aspects of incoming calls, as well as the not negligible differences in complexity proper between the 69 services, the figure is even more impressive. The one contrast that approached statistical significance in the *Same*/*Different* dimension is most likely the result of an inevitable inherent arbitrariness with regard to *Ignored* classifications, but do not influence the *Successful* figure directly since *Ignored* calls by definition are ignored.

Consequently, it would seem that the Swedish Nuance speech recognizer is in fact enjoying a reasonably "sweet sixteen" in the real world.

## Acknowledgements

Thanks to former and current colleagues at Telia Research, SRI International and Voice Provider, too numerous to be listed here. Special thanks to of Bengt Persson, CEO at Voice Provider. Also thanks to Per Näsman, Jens Edlund and Åsa Wengelin for comments on the statistical analysis.

## References

Carletta, J. (1996) Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22(2):249–254.

Cicchetti, D. V. & A. R. Feinstein (1990) High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* 43(6):551–558.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Di Eugenio, B. & M. Glass (2004) The Kappa Statistic: A Second Look. *Computational Linguistics* 30(1):95–101.

Eklund, R., J. Kaja, L. Neumeyer, F. Weng & V. Digalakis (2000) Porting a Recognizer to a New Language. In: M. Rayner et al. (2000), chapter 17, 265–273.

Feinstein, A. R. & D. V. Cicchetti (1990) High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43(6):543–549.

Frøkjær, E., M. Hertzum & K. Hornbæk (2000) Measuring Usability; Are Effectiveness, Efficiency, and Satisfaction Really Correlated? *CHI Letters* 2:345–352.

Hura, S. L. (2008) Voice User Interfaces. In: P. Kortum (ed.), *HCI Beyond the GUI. Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. Burlington, MA: Morgan Kaufmann.

Hornbaek, K. & E. L.-C. Law. (2007) Meta-Analysis of Correlations Among Usability Measures. *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 617–626.

Hornbaek, K. (2006) Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human–Computer Studies* 64:79–102.

Krippendorff, K. (2004) Reliability in Content Analysis. Some Common Misconceptions and Recommendations. *Human Communication Research* 30(3):411–433.

Rayner, M., D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.) (2000) *The Spoken Language Translator*, Cambridge: Cambridge University Press.

Walker, M. A., J. Fromer, G. Di Fabbrizio, C. Mestel & D. Hindle (1998) What Can I Say? Evaluating a Spoken Language Interface to Email. *CHI 98*, 582–589.

Proceedings

# FONETIK 2012

The XXVth Swedish Phonetics Conference
May 30–June 1, 2012

UNIVERSITY OF GOTHENBURG
PHILOSOPHY, LINGUISTICS & THEORY OF SCIENCE

Proceedings FONETIK 2012
The XXV$^{th}$ Swedish Phonetics Conference,
held at University of Gothenburg, May 30–June 1, 2012
Edited by Anders Eriksson and Åsa Abelin
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Box 200, SE 405 30 Gothenburg

ISBN 978-91-637-0985-2

Printed by Reprocentralen, Humanisten, University of Gothenburg.