

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,
Petitioner,

v.

BUFFALO PATENTS LLC,
Patent Owner.

Case No. IPR2023-01386
U.S. Patent No. 6,904,405

**DECLARATION OF CHRISTOPHER SCHMANDT IN SUPPORT OF
PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 6,904,405**

TABLE OF CONTENTS

	Page
I. INTRODUCTION	1
II. QUALIFICATIONS	2
III. MATERIALS CONSIDERED	6
IV. LEGAL STANDARDS	8
V. LEVEL OF ORDINARY SKILL IN THE ART	11
VI. CLAIM CONSTRUCTION	12
VII. OVERVIEW OF THE '405 patent.....	14
1. Prosecution History.....	16
VIII. Overview of Prior Art.....	18
A. Overview of <i>Waibel</i>	18
1. <i>Waibel's</i> Incorporation of <i>Suhm</i> , <i>Vo</i> , and <i>Manke</i>	21
a. Overview of <i>Suhm</i>	22
b. Overview of <i>Vo</i>	24
c. Overview of <i>Manke</i>	28
B. Overview of <i>Roberts</i>	29
C. Overview of <i>Verrier</i>	31
D. Overview of <i>Roth</i>	32
IX. Claims 1-12 Are Unpatentable	33
A. Claims 1-3, 6-9, and 12 Would Have Been Obvious over <i>Waibel</i> (incorporating <i>Suhm</i> , <i>Vo</i> , and <i>Manke</i>) and <i>Roberts</i> (Ground 1) or over <i>Waibel</i> in view of <i>Suhm</i> , <i>Vo</i> , <i>Manke</i> , and <i>Roberts</i> (Ground 2).....	33
1. Independent Claim 1	34

a.	[1P] “ <i>A system for generating text responsive to message input of a first type and a second type, the system comprising:</i> ”	34
b.	[1a] “ <i>(a) a unified message model including: (1) a shared language model; (2) a first model specific to a first type of message recognition; and (3) a second model specific to a second type of message recognition;</i> ”	35
i.	“ <i>(1) a shared language model</i> ”	36
ii.	“ <i>(2) a first model specific to a first type of message recognition</i> ”	43
iii.	“ <i>(3) a second model specific to a second type of message recognition</i> ”	44
c.	[1b] “ <i>(b) a first message recognizer, responsive to the first type of message input to provide text data in accordance with both the shared language model and the first model; and</i> ”	46
d.	[1c] “ <i>(c) a second message recognizer, responsive to the second type of message input to provide text data in accordance with both the shared language model and the second model</i> ”	47
e.	[1d] “ <i>(d) the shared language model is trainable responsive to user correction of misrecognition by either of the first and second message recognizers, thereby improving accuracy of each of the first and second message recognizers</i> ”	48
2.	Claim 2	57
3.	Claim 3	58
4.	Claim 6	61
5.	Independent Claim 7	63

a.	[7P] “ <i>A method for performing message recognition with a shared language model, the method comprising:</i> ”	63
b.	[7a] “ <i>(a) performing message recognition of a first type, responsive to a first type of message input, to provide text data in accordance with both the shared language model and a first model specific to the first type of message recognition;</i> ”	63
c.	[7b] “ <i>(b) performing message recognition of a second type, responsive to a second type of message input, to provide text data in accordance with both the shared language model and a second model specific to the second type of message recognition; and</i> ”	63
d.	[7c] “ <i>(c) training the shared language model responsive to user correction of error in message recognition of either of the first and second types, thereby improving accuracy of each of the first [and] second types of message recognition.</i> ”	64
6.	Claim 8	64
7.	Claim 9	65
8.	Claim 12	65
B.	Claims 4, 5, 10, and 11 Would Have Been Obvious over <i>Waibel, Roberts, and Verrier</i> (Ground 3) or over <i>Waibel, Suhm, Vo, Manke, Roberts, and Verrier</i> (Ground 4).....	66
1.	Claim 4	66
2.	Claim 5	70
3.	Claim 10	72
4.	Claim 11	72
C.	Claims 1-3 and 7-9 Would Have Been Obvious over <i>Roth and Roberts</i> (Ground 5).....	73

1.	Independent Claim 1	73
a.	[1P] “ <i>A system for generating text responsive to message input of a first type and a second type, the system comprising:</i> ”	73
b.	[1a] “ <i>(a) a unified message model including: (1) a shared language model; (2) a first model specific to a first type of message recognition; and (3) a second model specific to a second type of message recognition;</i> ”	74
i.	(1) a shared language model.....	75
ii.	(2) a first model specific to a first type of message recognition	76
iii.	(3) a second model specific to a second type of message recognition.....	77
c.	[1b] “ <i>(b) a first message recognizer, responsive to the first type of message input to provide text data in accordance with both the shared language model and the first model; and</i> ”	78
d.	[1c] “ <i>(c) a second message recognizer, responsive to the second type of message input to provide text data in accordance with both the shared language model and the second model</i> ”	78
e.	[1d] “ <i>(d) the shared language model is trainable responsive to user correction of misrecognition by either of the first and second message recognizers, thereby improving accuracy of each of the first and second message recognizers</i> ”	79
2.	Claim 2	87
3.	Claim 3	89
4.	Independent Claim 7	91

a.	[7P] “ <i>A method for performing message recognition with a shared language model, the method comprising:</i> ”	91
b.	[7a] “ <i>(a) performing message recognition of a first type, responsive to a first type of message input, to provide text data in accordance with both the shared language model and a first model specific to the first type of message recognition;</i> ”	92
c.	[7b] “ <i>(b) performing message recognition of a second type, responsive to a second type of message input, to provide text data in accordance with both the shared language model and a second model specific to the second type of message recognition; and</i> ”	92
d.	[7c] “ <i>(c) training the shared language model responsive to user correction of error in message recognition of either of the first and second types, thereby improving accuracy of each of the first and [sic] second types of message recognition.</i> ”	92
5.	Claim 8	93
6.	Claim 9	93
D.	Claims 4, 5, 10, and 11 Would Have Been Obvious over <i>Roth</i> , <i>Roberts</i> , and <i>Verrier</i> (Ground 6)	94
1.	Claim 4	94
2.	Claim 5	96
3.	Claim 10	98
4.	Claim 11	98
E.	Claims 6 and 12 Would Have Been Obvious over <i>Roth</i> , <i>Roberts</i> , and <i>Vo</i> (Ground 7)	98
1.	Claim 6	98

2.	Claim 12.....	100
X.	Conclusion	100

CLAIMS 1-12 OF THE '405 PATENT:

1. A system for generating text responsive to message input of a first type and a second type, the system comprising:

(a) a unified message model including:

- (1) a shared language model;
- (2) a first model specific to a first type of message recognition; and
- (3) a second model specific to a second type of message recognition;

(b) a first message recognizer, responsive to the first type of message input to provide text data in accordance with both the shared language model and the first model; and

(c) a second message recognizer, responsive to the second type of message input to provide text data in accordance with both the shared language model and the second model; wherein

(d) the shared language model is trainable responsive to user correction of misrecognition by either of the first and second message recognizers, thereby improving accuracy of each of the first and second message recognizers.

2. The system of claim 1 wherein:

(a) the first type of message recognition is speech recognition, the first type of message input is voice input, and the first model is an acoustic model; and

(b) the second type of message recognition is handwriting recognition, the second type of message input is freehand input, and the second model is a handwriting model.

3. The system of claim 2 further comprising a stylus, wherein the shared language model is trainable responsive to user correction of misrecognition by the first message recognizer by (a) movement of the stylus across visual indicia of dictated text that has been erroneously generated to specify a selection from the text, and (b) correction of the selection.

4. The system of claim 3 wherein:

(a) the stylus includes a pointing interface and a microphone;

(b) the system further comprises a tablet responsive to the pointing interface;

(c) the voice input is from the microphone; and

(d) the freehand input is from the pointing interface being manually manipulated on the tablet.

5. The system of claim 4 wherein the visual indicia of dictated text appears on the tablet.

6. The system of claim 1 wherein the shared language model comprises a syntactic model and a semantic model.

7. A method for performing message recognition with a shared language model, the method comprising:

(a) performing message recognition of a first type, responsive to a first type of message input, to provide text data in accordance with both the shared language model and a first model specific to the first type of message recognition;

(b) performing message recognition of a second type, responsive to a second type of message input, to provide text data in accordance with both the shared language model and a second model specific to the second type of message recognition; and

(c) training the shared language model responsive to user correction of error in message recognition of either of the first and second types, thereby improving accuracy of each of the first and second types of message recognition.

8. The method of claim 7 wherein:

(a) performing the first type of message recognition comprises performing speech recognition responsive to voice input in accordance with the shared language model and a first model that is an acoustic model; and

(b) performing the second type of message recognition comprises performing handwriting recognition responsive to freehand input in accordance with the shared language model and a second model that is a handwriting model.

9. The method of claim 8 wherein training the shared language model includes (a) moving a stylus across visual indicia of dictated text that has been erroneously generated to specify a selection from the text, and (b) correcting the selection.

10. The method of claim 9 wherein:

(a) performing speech recognition responsive to voice input includes responding to voice input from a microphone in the stylus; and

(b) performing handwriting recognition responsive to freehand input includes responding to manual manipulation of the stylus on a tablet.

11. The method of claim 10 wherein the visual indicia of dictated text appears on the tablet.

12. The method of claim 7 wherein performing message recognition in accordance with the shared language model comprises performing message recognition in accordance with a syntactic model and a semantic model.

I. INTRODUCTION

1. I, Christopher Schmandt, submit this declaration to state my opinions on the matters described below.

2. I have been retained on behalf of Google LLC to serve as an expert in this proceeding before the United States Patent & Trademark Office.

3. I am over 21 years of age and, if I am called upon to do so, I would be competent to testify as to the matters set forth herein.

4. My compensation is in no way contingent on the nature of my findings, the presentation of my findings in testimony, or the outcome of any proceeding.

5. I understand that this proceeding involves U.S. Patent No. 6,904,405 which is entitled “Message Recognition Using Shared Language Model” (“the ’405 patent”), and I have been asked to provide my opinions as to the patentability or unpatentability of the claims of the ’405 patent.

6. I understand the ’405 patent has been provided as Ex. 1001. I have reviewed and am familiar with the ’405 patent. I understand that the file history has been provided as Ex. 1002. I have reviewed and am familiar with the file history of the ’405 patent.

7. This declaration sets forth my opinions, which I formed based on my study of the evidence; my understanding as an expert in the field; and my training, education, research, knowledge, and personal and professional experience.

8. Throughout this declaration, I refer to specific pages, figures, and/or line numbers of various exhibits. These citations are illustrative and are not intended to suggest that they are the only support for the propositions for which they are cited.

II. QUALIFICATIONS

9. I retired four years ago after a 40-year career at the Massachusetts Institute of Technology (MIT); for most of that time, I was employed as a Principal Research Scientist at the Media Laboratory. In that role, I also served as faculty for the MIT Media Arts and Sciences academic program. I was a founder of the Media Laboratory, a research lab which now spans two buildings.

10. I received my B.S. degree in Electrical Engineering and Computer Science from MIT in 1978, and my M.S. in Visual Studies (Computer Graphics), also from MIT, the following year. I was employed at MIT starting in 1980, initially at the Architecture Machine Group which was an early computer graphics and interactive systems research Lab. In 1985, I helped found the Media Laboratory and continued to work there until retirement. I was director of a research group titled "Living Mobile." My research spanned distributed

communication and collaborative systems, with an emphasis on multimedia and user interfaces, with a strong focus on speech-based systems and portable devices. I have over 70 published conference and journal papers and one book in the field of speech technology and user interaction.

11. For the first fifteen years of my career, my research emphasized speech recognition and speech user interfaces. I built the first system utilizing multimodal user input—speech and gesture—starting in 1980 (“Put That There”); this remains known as a seminal work in the field. “Put That There” was also the first conversational computer system utilizing speech recognition and synthesis to engage in a voice-user interaction. I continued to innovate speech user interfaces using recognition, text-to-speech synthesis, and recorded audio in a wide variety of projects. I built one of the first graphical user interfaces for audio editing, employing text keyword recognition on voice memos in 1982 (Intelligent Ear). I built the first research-grade unified messaging system, which combined text and voice messages into a single inbox, with speech recognition over the phone for remote access, and a graphical user interface for desktop access in 1983 (Phone Slave). I built a speech recognizer to detect acoustic correlates of mouth configuration to drive a real time animated character (Lip Sync, 1983). Along with my students, we built the first system for real-time spoken driving directions, including speech-accessible maps of Cambridge, Massachusetts in 1987 (Back

Seat Driver). We built some of the earliest speech-based personal assistants for managing messages, calendars, contacts, etc. (Conversational Desktop 1985, Chatter 1993, MailCall 1996). We applied speech recognition to large bodies of everyday conversations captured with a wearable device and utilized as a memory aid (Memory Prosthesis 2004). We used speech recognition on radio newscasts to build a personalized version of audio newscasts (Synthetic News Radio, 1999) and also investigated adding speech recognition to a mouse-based window system a few years earlier.

12. With the appearance of handheld devices, some of which were capable of recognizing handwriting, I used and built applications for these. The first was the Apple Newton, for which we built a voice-based reminder system with multiple “to do” categories and time-compressed speech playback (VoiceNotes, 1992). We built quite a few systems employing speech recognition in handheld mobile devices (ComMotion 1999, Nomadic Radio 2000, Impromptu 2001, and Symphony 2004, for example). We applied speech recognition to large bodies of everyday conversations captured with a wearable device and utilized as a memory aid (Memory Prosthesis 2004). In this context we worked with devices such as the Palm Pilot, which used Graffiti for handwriting recognition, the HP Ipaq handheld computer, pagers, flip phones, and finally smart phones.

13. I was later awarded the prestigious Association for Computing Machinery (ACM) Computer Human Interface (CHI) Academy membership specifically for those years of work pioneering speech user interfaces.

14. In the course of my research I built a number of speech recognition client/server distributed systems, with the first being in 1985. Much of the initial motivation for a server architecture was that speech recognition required expensive digital signal processing hardware that we could not afford to put on each computer, so a central server with the required hardware was used. Later versions of the speech recognition server architecture allowed certain computers to perform specialized tasks serving a number of client computers providing voice user interfaces, either on screens or over telephone connections.

15. Because of my early work with distributed speech systems, I served for several years in the mid-1990s with a working group on the impact of multimedia systems on the Internet reporting to the Internet Engineering Task Force (IETF) and later the Internet Activities Board (IAB). This work impacted emerging standards such as Session Initiation Protocol (SIP).

16. In my faculty position I taught graduate-level courses in speech technology and user interaction design, and directly supervised student research and theses at the Bachelors, Masters, and PhD level. I oversaw the Masters and PhD thesis programs for the entire Media Arts and Sciences academic program

during my more senior years. I also served on the Media Laboratory Intellectual Property Committee for many years, including a period during which I was the chair.

III. MATERIALS CONSIDERED

17. In forming my opinions, I have reviewed the following documents, and any other document cited in this declaration:

Exhibit	Description
Ex. 1001	U.S. Patent No. 6,904,405 to Suominen (“the ’405 patent”)
Ex. 1002	Prosecution History of U.S. Patent No. 6,904,405
Ex. 1003	Curriculum Vitae of Christopher Schmandt
Ex. 1005	A. Waibel et al., “Multimodal Interfaces for Multimedia Information Agents,” IEEE (1997) (“ <i>Waibel</i> ”)
Ex. 1006	WO 92/05517 to Roth (“ <i>Roth</i> ”)
Ex. 1007	U.S. Patent No. 5,027,406 to Roberts et al. (“ <i>Roberts</i> ”)
Ex. 1008	B. Suhm et al., “JANUS: Towards Multilingual Spoken Language Translation,” Proc. ARPA SLT Workshop 95 (Austin, Texas) (“ <i>Suhm</i> ”)
Ex. 1009	Vo et al., “Building and Application Framework for Speech and Pen Input Integration in Multimodal Learning Interfaces,” Proc. ICASSP ’96 (Atlanta, GA) (“ <i>Vo</i> ”)
Ex. 1010	S. Manke et al., “The Use of Dynamic Writing Information in a Connectionist On-Line Cursive Handwriting Recognition System,” <i>Advances in Neural Information Processing Systems</i> 6, Morgan Kaufmann, 1994 (“ <i>Manke</i> ”)
Ex. 1011	EP 0622724 to Verrier (“ <i>Verrier</i> ”)

Exhibit	Description
Ex. 1012	L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Procs. of the IEEE, 77, 257 (1989) (“ <i>Rabiner</i> ”)
Ex. 1013	Klevans et al., “Voice Recognition,” 38-45, 1997 (“ <i>Klevans</i> ”)
Ex. 1014	F. Jelinek, “Statistical Methods for Speech Recognition,” 39-54, 3d ed., 2001 4 (“ <i>Jelinek</i> ”)
Ex. 1015	Rabiner et al., “Fundamentals of Speech Recognition,” 321-386, 1993 (“ <i>Rabiner & Juang</i> ”)
Ex. 1016	Bellegarda et al. “A Fast Statistical Mixture Algorithm for On-Line Handwriting Recognition,” 1227-1233, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994
Ex. 1017	U.S. Patent No. 5,459,798 to Bailey et al. (“ <i>Bailey</i> ”)
Ex. 1018	U.S. Patent No. 5,855,000 to Waibel et al. (“’000 Patent”)
Ex. 1023	Starner et al. “On-Line Cursive Handwriting Recognition Using Speech Recognition Methods” Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing, 1994 (“ <i>Starner</i> ”)
Ex. 1026	Bellegarda, et al., “A Novel Word Clustering Algorithm Based on Latent Semantic Analysis,” IEEE, 1996 (“ <i>Bellegarda</i> ”)
Ex. 1029	Moran et al., “Multimodal User Interfaces in the Open Agent Architecture,” Knowledge-Based Systems Vol. 10 295-303, 1998 (“ <i>Moran</i> ”)

18. I have also relied on my education, experience, research, training, and knowledge in the relevant art, and my understanding of any applicable legal principles described in this declaration.

19. All of the opinions contained in this declaration are based on the documents I reviewed and my knowledge and professional judgment. My opinions have also been guided by my understanding of how a person of ordinary skill in the art would have understood the claims of the '405 patent at the time of the alleged invention. For purposes of this declaration, I have been asked to assume that the date of the alleged invention is the earliest claimed priority date: July 17, 1999.

20. I reserve the right to supplement and amend any of my opinions in this declaration based on documents, testimony, and other information that becomes available to me after the date of this declaration.

IV. LEGAL STANDARDS

21. For purposes of this declaration, I have been asked to opine only on issues regarding obviousness under 35 U.S.C. § 103. I have been informed of the following legal standards, which I have applied in forming my opinions.

22. I have been asked to provide my opinions as to whether the cited prior art teaches or renders obvious claims 1-12 of the '405 patent from the perspective of a person of ordinary skill in the art at the '405 patent's priority date in 1999, as described in more detail below.

23. For purposes of this declaration, I have been informed and understand certain aspects of the law as it relates to my opinions.

24. I have been advised and understand that there are two ways in which prior art may render a patent claim unpatentable. First, I have been advised that the prior art can “anticipate” a claim. Second, I have been advised that the prior art can make a claim “obvious” to a person of ordinary skill in the art. I understand that for an invention claimed in a patent to be patentable, it must not be anticipated and must not be obvious based on what was known before the invention was made.

25. I have been advised and understand the information used to evaluate whether an invention was new and not obvious when made is generally referred to as “prior art.” I understand that prior art includes patents and printed publications that existed before the earliest filing date of the patent (which I have been informed is also called the “effective filing date”). I have been informed and understand that a patent or published patent application is prior art if it was filed before the earliest filing date of the claimed invention and that a printed publication is prior art if it was publicly available before the earliest filing date.

26. I have been advised and understand that a dependent claim is a patent claim that refers back to another patent claim. I have been informed and understand that a dependent claim includes all of the limitations of the claim to which it refers.

27. I have been advised and understand that a patent claim may be invalid as obvious under 35 U.S.C. § 103 if the differences between the subject matter

claimed and the prior art are such that the claimed subject matter as a whole would have been obvious to a person of ordinary skill in the art at the time the invention was made. I have also been advised that several factual inquiries underlie a determination of obviousness. These inquiries include (1) the scope and content of the prior art, (2) the level of ordinary skill in the field of the invention, (3) the differences between the claimed invention and the prior art, and (4) any objective evidence of non-obviousness (which I have been informed may also be called “secondary considerations”).

28. I have also been advised and understand that, when obviousness is based on a combination of references, that party must identify a reason why a person skilled in the art would have been motivated to combine or modify the asserted references in the manner recited in the claims and to explain why one skilled in the art would have had a reasonable expectation of success in making such combinations or modifications.

29. I have been advised and understand that the law permits the application of “common sense” in examining whether a claimed invention would have been obvious to a person skilled in the art. For example, I have been advised that combining familiar elements according to known methods and in a predictable way may suggest obviousness when such a combination would yield nothing more than predictable results. I understand, however, that a claim is not obvious merely

because every claim element is disclosed in the prior art and that a party asserting obviousness must still provide a specific motivation to combine or modify the references as recited in the claims and explain why one skilled in the art would have reasonably expected to succeed in doing so.

30. I have been advised and understand that two references are considered to be in the same field of art when the references are either (1) in the same field of endeavor, regardless of the problems they address, or (2) reasonably pertinent to the particular problem being solved by the inventor in his patent.

31. I am not aware of any evidence of secondary considerations that would support a determination of non-obviousness of the claimed subject matter in the '405 patent.

32. I have been informed that in *inter partes* review proceedings, such as this one, the party challenging the patent bears the burden of proving unpatentability by a preponderance of the evidence. I understand that a preponderance of the evidence means “more likely than not.”

V. LEVEL OF ORDINARY SKILL IN THE ART

33. In rendering the opinions set forth in this declaration, I have been asked to consider the '405 patent's claims and the prior art through the eyes of a person of ordinary skill in the art (which I may also refer to as “one skilled in the art,” “skilled artisan,” “POSITA,” or similar variation). I have considered factors

such as the educational level and years of experience of those working in the pertinent art, the types of problems encountered in the art, the teachings of the prior art, patents and publications of other persons or companies, and the sophistication of the technology.

34. I have been instructed to assume a person of ordinary skill in the art is not a specific real individual, but rather a hypothetical individual having the qualities reflected by the factors discussed above.

35. Taking these factors into consideration, it is my opinion that a person of ordinary skill in the art of the '405 patent as of its filing date, would have at least a bachelor's degree in computer engineering, computer science, electrical engineering, or a similar field, and approximately three years of industry or academic experience in a field related to acoustics, speech recognition, speech detection, handwriting recognition, or signal processing. Work experience can substitute for formal education and additional formal education can substitute for work experience.

36. By July 17, 1999, I was at least a person of ordinary skill in the art, as shown by my qualifications and work experience above.

37. In this declaration, and for all of my opinions herein, I have applied the knowledge of a person of ordinary skill in the art as of July 17, 1999.

VI. CLAIM CONSTRUCTION

38. I have been instructed that the words of a claim are typically given their ordinary and customary meaning, as they would have been understood by a person of ordinary skill in the art at the time of the invention, in view of the intrinsic record (discussed below). In this case, I have been instructed to assume that the “time of the invention” for purposes of claim construction is July 17, 1999, which is the earliest claimed priority filing date for the ’405 patent. The opinions herein pertain to that time frame, except where expressly stated otherwise.

39. I have been instructed that the “intrinsic record” includes the patent itself, including the claims, description, and figures (Ex. 1001), and the patent’s prosecution history—i.e., the record of proceedings at the U.S. Patent and Trademark Office (“Patent Office”) concerning the patent (Ex. 1002). I understand that, like the claims and written description, the prosecution history provides evidence to a person of ordinary skill in the art of how the inventor intended his patent to be understood, and how the Patent Office understood the patent. I understand that the inventor is permitted to apply a special definition to the terms or to limit the scope of claim terms in his patent claims, which may differ from the term’s plain and ordinary meaning. That special definition or limitation on scope may be provided in the patent’s written description, the patent’s prosecution history, or both.

40. I understand that claim interpretation may also be informed by “extrinsic evidence” (that is, evidence outside of the patent record itself). I have been informed that extrinsic evidence may include dictionaries, technical treatises, and other materials evidencing the meaning of a claim term and the understanding held by a POSITA in the relevant time period.

41. I have been asked for purposes of this declaration to apply the plain and ordinary meaning of the claim terms as they would have been understood by a person of ordinary skill in the art at the earliest claimed priority date of the ’405 patent (July 17, 1999).

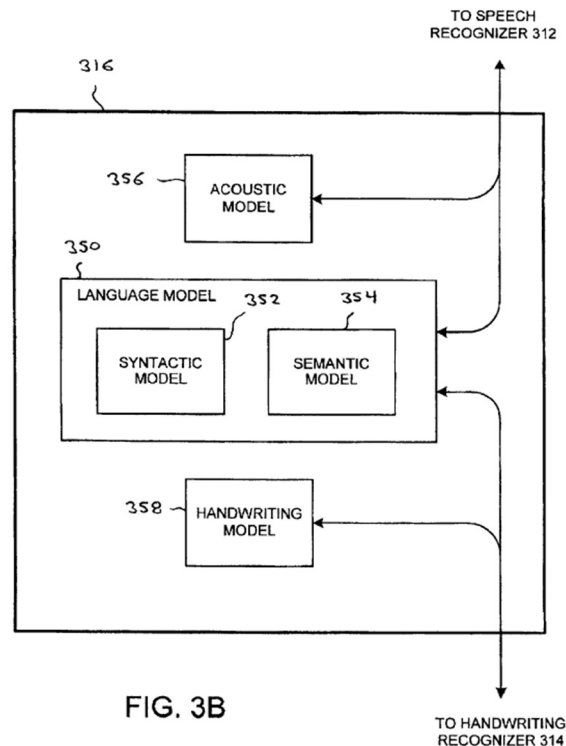
42. I understand that no claim term’s construction is in dispute at this time. Based on my review of these materials and my personal knowledge and experience, I have considered each term of the ’405 patent as it would have been understood by one skilled in the art as of July 17, 1999.

VII. OVERVIEW OF THE ’405 patent

43. The ’405 patent discloses “methods and systems [that] perform multiple different types of message recognition using a shared language model.” ’405 patent, Abstract.

44. I understand that the ’405 patent explains a shared language model is “any language model that may be shared between two types of subordinate message recognizers included within a message recognizer.” *Id.*, 22:46-59. In an

exemplary embodiment, the specification discloses an acoustic model 356 for use by a speech recognizer, a handwriting model 358 for use by a handwriting recognizer, and a language model 350 shared between the speech recognizer and the handwriting recognizer. *Id.*, 22:52-23:8 Fig. 3B. These are shown below in Figure 3B.



Id., Fig. 3B; *see also, e.g., id.*, 22:60-23:16. According to the '405 patent, the speech recognizer provides text data in response to speech input using the acoustic model and the shared language model, and the handwriting recognizer provides text data in response to handwriting input using the handwriting model and the shared language model. *Id.*, 22:65-23:2, Fig. 3B. The '405 patent explains the shared language model can be trained by a user making correction to any

misrecognition by either the speech or the handwriting recognizer. *Id.*, 23:2-8, claim 1.

1. Prosecution History

45. I was informed that, during prosecution of the '405 patent, another prior-art reference attributed to *Waibel*, U.S. Patent No. 5,855,000, was considered by the Examiner during prosecution. I believe that the '000 patent describes a generalized system, which is not the same as the specific system disclosed in *Waibel*.

Regardless, Applicant's comments distinguishing the '000 Patent over its claims were misleading at best. For example, Applicant claimed it "does not believe that *Waibel* teaches or suggests any type of language model ... shared between the recognizers for the primary and secondary input signals ... " Ex. 1002 at 55-56.

But as relevant to the claims, a POSITA would have understood the experiments in the '000 *Waibel* patent included different language models shared between recognizers for the primary and secondary input signals. *See, e.g.*, '000 patent, 15:43-51, 15:61-64, 6:28-65, 7:19-26. For example, a POSITA would have understood that the purpose of the language models in the '000 patent is to facilitate recognition, and in a system where inputs, including speech and handwriting, are considered primary or secondary inputs, *see, e.g.*, '000 patent, 5:39-41, it would have been a poor design to use the language model for one input modality and not the other. In fact, when recognition is constrained to vocabulary

and grammar of a particular task, a POSITA would have found it counterproductive to apply one model to one recognizer, but a different model to another.

VIII. Overview of Prior Art

A. Overview of *Waibel*

46. According to *Waibel*, existing human-computer interfaces were “limited and inflexible,” failing to take advantage “of the many communication channels humans use to communicate [in] verbal and non-verbal ways.” *Waibel*, 167. To address this, and to build computer systems “that are more intuitive” and that “take advantage of a rich spectrum of cues” (like verbal and non-verbal), *Waibel* describes a multimodal system and method of generating text. *Waibel*, Abstract. Specifically, *Waibel* teaches generating text jointly in response to inputs from multiple modalities (such as speech and pen input). *Waibel*, Abstract, 167.

47. Some of *Waibel*’s multimodal components and inputs include:

48. 1) A speech recognition subsystem “based on the recognition front-end of the JANUS speech translation system [5] which is capable of processing speaker-independent, spontaneous speech” that “can be adapted to any task domain by retraining the language models and possibly tuning the acoustic models if the domain involves special vocabulary.” *Id.*, 167. *Waibel* lists incorporated reference [5] as

- [5] B. Suhm, P. Geutner, T. Kemp, A. Lavie, L.J. Mayfield, A. McNair, I. Rogina, T. Schultz, T. Sloboda, W. Ward, M. Woszczyna and A. Waibel: “JANUS: Towards Multilingual Spoken Language Translation,” *Proc. ARPA SLT Workshop 95* (Austin, Texas).

Id., 170. Incorporated reference [5] is included herein as Ex. 1008.

49. 2) A “MS-TDNN-based handwriting recognizer [7] is capable of processing writer-independent, continuous (cursive) handwriting,” which can “invoke handwriting recognition on pen input.” *Id.*, 167. *Waibel* lists incorporated reference [7] as

[7] S. Manke, M. Finke and A. Waibel: “The Use of Dynamic Writing Information in a Connectionist On-Line Cursive Handwriting Recognition System,” *Advances in Neural Information Processing Systems 6*, Morgan Kaufmann, 1994.

Id., 170. Incorporated reference [7] is included herein as Ex. 1010.

50. I understand that *Waibel* discloses that the output from the processors of its different recognizers, including the above speech and handwriting recognizers, are input into a multimodal interpreter “capable of producing an interpretation of user intent ... from the output of the modality processors.” *Id.*, 167-68. *Waibel* states this is “joint interpretation.” *Id.* Specifically, *Waibel* explains:

[i]n our joint interpretation scheme, the user intent is represented by a frame consisting of slots specifying pieces of information such as the action to carry out or the parameters for that action. Recognition output from the modality processors are parsed into partially filled frames that are merged together to produce the combined interpretation as described in [6]. This technique leads to uniform handling of high-level information from all input sources, which is very important for modularity and

extensibility. To add another input modality we need only provide a module to convert low-level recognizer output to a partially filled frame to be merged with others. In addition, context information can be retained across input events by merging with previous interpretation frames.

Id., 168. *Waibel* lists incorporated reference [6] as

[6] M.T. Vo and C. Wood: "Building and application framework for speech and pen input integration in multimodal learning interfaces," *Proc. ICASSP'96* (Atlanta, GA).

Id., 170. Incorporated reference [6] is included herein as Ex. 1009.

51. *Waibel* further discloses Multimodal Error Correction. *Id.*, 168.

According to *Waibel*, the potential for error in speech interfaces can "be balanced by the redundant and alternate ability to enter information or queries by spelled, handwritten or typed input." *Id.* As *Waibel* explains:

Our approach is to have the user collaborate with the system in recovering from interpretation errors [10]. First, errors have to be identified, either initiated by the system, for instance based on some confidence measure, or by the user. If a graphical user interface (GUI) is available, the user can simply highlight erroneous words (the reparandum) in the recognition hypothesis displayed. For error correction, the user provides additional input, choosing among different correction methods: repair by repeating the reparandum by respeaking, spelling out

loud, or handwriting; repair by paraphrasing the reparandum; repair by pen gestures (e.g. to delete or insert), in addition to the standard repair by typing or selecting among N-best alternatives.

Id., 168.

52. Moreover,

[w]e have built prototypical speech user interfaces with error repair capabilities for dictation and form filling tasks. In this application context, requiring the user to identify errors by highlighting them is appropriate, since the user naturally focuses on the actual recognition hypothesis. For error correction, we offer to repeat the input (potentially switching to another modality), and pen gestures to indicate where to insert words, or which word to delete.

Id.

1. *Waibel's Incorporation of Suhm, Vo, and Manke*

53. As shown above, in addition to its broad teachings, *Waibel* refers to and incorporates three papers, *Suhm*, *Vo*, and *Manke*. *Id.*, 167-68, 170. Of note, *Suhm*, *Vo*, *Manke*, and *Waibel* arise from the same laboratory in the same university, from around the same time period, and share a number of the same authors. It was common practice in academia to publish separate papers each describing a component of a single multi-part system.

54. In my opinion, *Suhm*, *Vo*, and *Manke* are clearly identified in *Waibel*, and explicitly stated as respectively describing its speech recognizer, joint recognizer, and handwriting recognizer. *Waibel*, 167-68, 170. Because *Waibel* identifies the papers (and not just particular portions of their disclosures) and where they can be found (*id.*, 170), I understand that the papers are incorporated in their entirety.

a. Overview of *Suhm*

55. *Suhm* describes the speech recognizer disclosed in *Waibel*. *Waibel*, 167, 170.¹ Specifically, *Suhm* describes the JANUS speech translation system, in which spoken input is translated into either English text output or another language output. *Waibel*, 167; *Suhm*, 221. According to *Suhm*, to “tackle the difficulty of spontaneous human-human dialogs[,] we improved the JANUS-2 recognizer along its three knowledge sources[:] acoustic models, dictionary and language models.” *Suhm*, 221. It explains that its “robust translation system” uses semantic analysis and is “suited to processing spontaneous speech.” *Id.* The system modules for the speech recognizer are “speech recognition, parsing, discourse processing, and generation.” *Id.* *Suhm* explains that each module is “language-independent”

¹ Two of the listed authors of *Suhm*, B. Suhm and A. Waibel, are also listed authors in *Waibel*. Compare *Suhm*, 221, with *Waibel*, 167.

because each “consists of a general processor that applies independently specified knowledge about different languages.” *Id.*, 221-22.

56. *Suhn* explains that its speech recognition module will decode speech from the source language “into a list of sentence candidates” represented “as a word lattice or Nbest list.” *Id.*, 222. “As [a] last step, the final language independent representation is sent to the generator to be translated in any of the target languages.” *Id.* The system architecture is shown below in Figure 1:

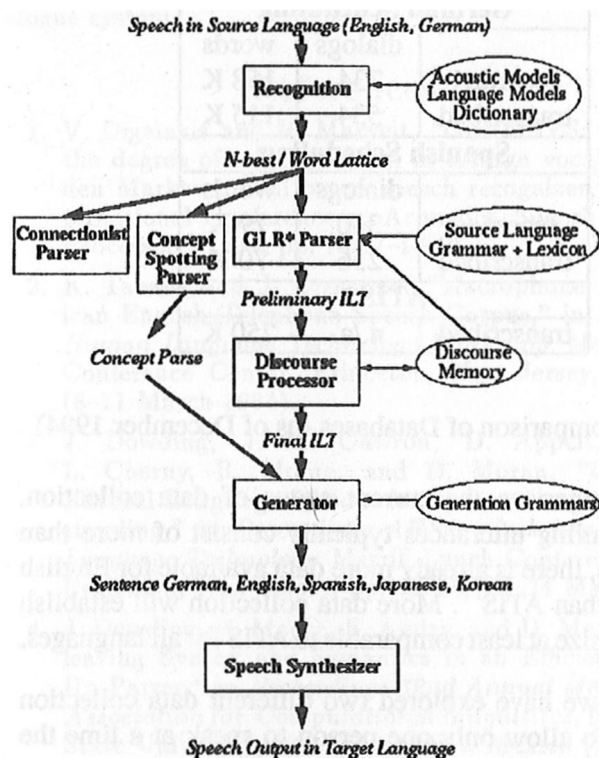


Figure 1: System Diagram

Id., 222 & Fig. 1. *Suhn* explains, after parsing, “a discourse processor can be used to put the current utterance in context” and permit the system to “dynamically

adapt the vocabulary and language model of the recognizer based on the current discourse state.” *Id.*, 222.

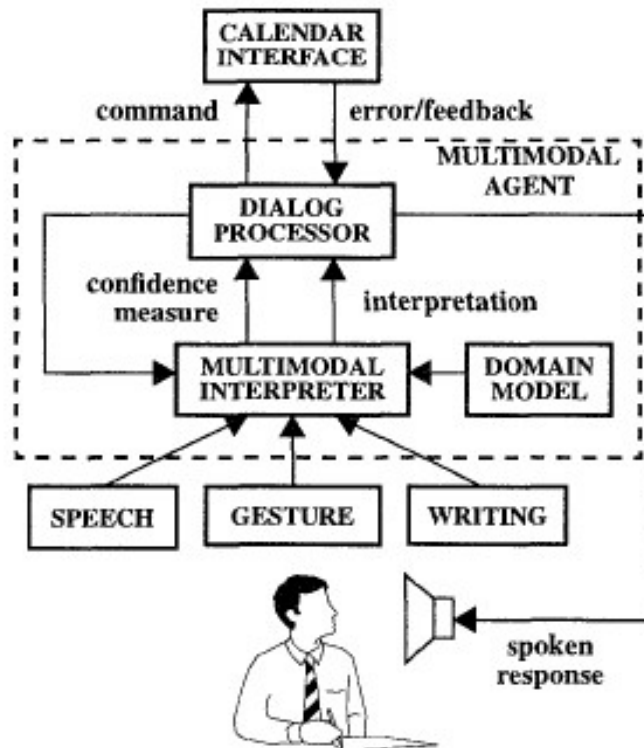
57. To improve the speech recognition component of its system, *Suhm* discloses a semi-continuous, phonetically tied Hidden Markov Model (HMM), and adapts the codebook size of each phoneme to the amount and distribution of training data. *Id.*, 223. “During training, the size of the codebook is incrementally increased.” *Id.* *Suhm* discloses a “standard word bigram language model” and a “data-driven approach to improve existing dictionaries and automatically add new words and pronunciation variants whenever needed.” *Id.*, 222-23. *Suhm* discloses that its dictionary “learning algorithm requires transcripts for the whole training set and a phoneme confusability matrix of the speech recognizer.” *Id.*, 223. It may also use morpheme-based language models to limit the growth rate for “morphologically rich languages.” *Id.*, 223-24.

b. Overview of *Vo*

58. *Vo* describes the joint interpreter disclosed in *Waibel*. See *Waibel*, 168, 170.² This joint interpreter jointly interprets speech, pen-based gestures, and handwriting information from different sources based on semantic frame merging.

² The first author, M. Vo, is also a listed author in *Waibel*. Compare *Vo*, 3545, with *Waibel*, 167.

See *Vo*, 3545. *Vo* explains that its joint “multimodal interpreter” produces an “interpretation of user intent” based on outputs from the modality processors. *Id.*, 3546. Specifically, these outputs are “parsed into partially filled frames that are merged together to produce the combined speech interpretation.” *Id.* Figure 1 is a block diagram of *Vo*’s system:



Id., 3545-46 & Fig. 1. According to *Vo*, the individual modality components (speech, gesture, and handwriting recognizers) may be separate modules and may be replaced with ease. *Id.*, 3545.³

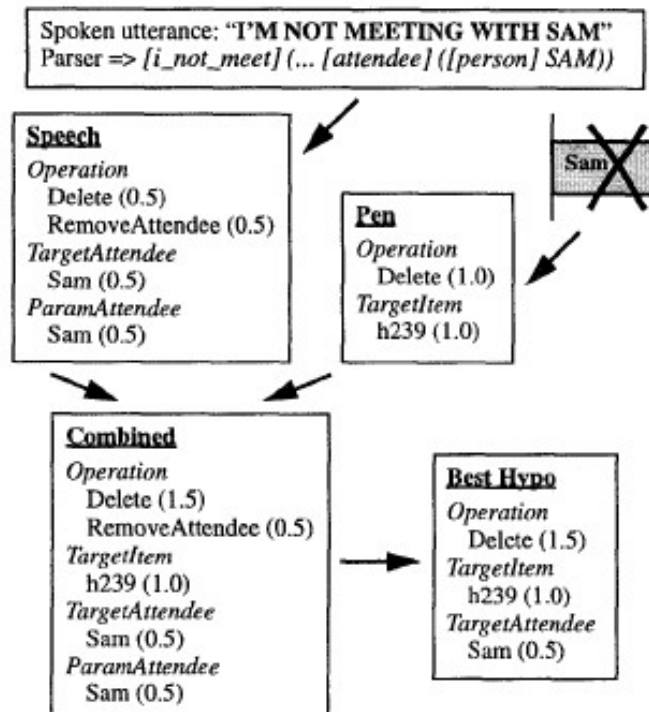
59. As to the speech recognizer, *Vo* explains that it may capture inputs by a speech recognizer. *Id.*, 3546. The resulting “text string output” from the speech recognizer is then processed by the Phoenix semantic parser.” *Id.* The parser “compiles a grammar specifying semantically meaningful fragments of text,” and this is compiled into a “network that identifies these fragments in the recognized utterance.” *Id.* The parser is “capable of skipping unknown words and unmatched fragments and can therefore deal gracefully with ungrammatical sentences.” *Id.*

60. As to the handwriting recognizer, *Vo* explains that it may capture inputs by a “touch-sensitive screen.” *Id.* It “produce[s] a sequence of gesture/handwriting components” that “may mean different things depending on the calendar context,” and the handwritten words are “encoded into frames according to simple hard-coded heuristics.” *Id.*

³ *Vo* also explains that its speech recognition subsystem is based on the JANUS speech translation system, which was discussed above with respect to *Suhm*. *Vo*, 3546. Its handwriting recognizer is based on the MS-TDNN. *Id.*

61. *Vo* explains that “[f]rames are merged by taking the union of the sets of values filling each slot[] and adding corresponding scores.” *Id.*, 3547. This “produces an aggregate frame encoding all alternative interpretation hypotheses.” *Id.* Figure 2 has one example of such “interpretation using frames”:

FIGURE 2: Example of interpretation using frames



Id., Fig. 2. *Vo* explains that the “multimodal interpreter extracts these hypotheses in descending order of aggregate scores and sends the best ones to the dialog processor” for processing. *Id.*, 3547. According to *Vo*, the “dialog processor maintains a single frame per dialog to hold all [the] information obtained since the start of the dialog; new information from each additional input event is merged into this frame and the scores are weighted appropriately so that frames produced by

different input events in the same dialog contribute equally to the end result.” *Id.*

The dialog processor will also “filter out hypotheses that would cause errors.” *Id.*

Once a “good hypothesis” is found, the dialog processor will send it to the “domain interface,” where a “user has the option of undoing the operation just performed if it is incorrect.” *Id.*

c. Overview of *Manke*

62. *Manke* describes the handwriting interpreter disclosed in *Waibel*. *Waibel*, 167, 170.⁴ *Manke* describes the NPen⁺⁺ system for writer-independent, large vocabulary, on-line cursive handwriting recognition. *Manke*, Abstract. Its system combines an input representation that “preserves the dynamic writing information” with “a neural network architecture,” MS-TDNN, which “integrates recognition and segmentation in a single framework.” *Id.*

63. During preprocessing, “a temporal sequence of *N*-dimensional feature vectors is computed from the original coordinate sequence” and recorded. *Id.*, 1094-95. The recognizer employs Hidden Markov Models (HMM) where “each character c_j itself is modelled by a three state hidden markov model.” *Id.*, 1096.

⁴ A. Waibel, who is a listed author in *Manke*, is the first author in *Waibel*. Compare *Manke*, 1093, with *Waibel*, 167.

64. Then, in the recognition processing stage, *Manke*'s system integrates word recognition and segmentation into a single network architecture using a model (MS-TDNN) that "was originally proposed for continuous speech recognition tasks." *Id.*, 1094. According to *Manke*, this approach "combines shift-invariant, high accuracy pattern recognition capabilities of a TDNN" with "a non-linear alignment procedure for aligning strokes into character sequences." *Id.*

65. *Manke* tested its model using the "CRT (Conference Registration Task) and WSJ (ARPA Wall Street Journal Task)" vocabularies, "originally defined for speech recognition evaluations." *Id.*, 1099. These vocabularies were chosen because of the "synergy effects between handwriting recognition and speech recognition, since ... the final goal is to integrate our speech recognizer" and the proposed NPen⁺⁺ system "into a multi-modal system." *Id.*

B. Overview of *Roberts*

66. *Roberts* discloses a speech recognition method and training models for use in speech recognition. *Roberts*, 1:6-11. It enables a user to "build acoustic models from tokens that are collected during actual use of the system rather than during a separate training or enrollment process," "easily correct errors made in continuous speech recognition," and "develop a language model for use in a speech recognition system" "without requiring much additional effort" to customize the language model for each user's "particular diction." *Id.*, 2:17-30.

67. *Roberts*'s training method includes first having the user "speak[] a word which may or may not already be in the active vocabulary." *Id.*, Abstract. The system then "displays a list of the words in the active vocabulary which best match the spoken word." *Id.* "By keyboard or voice command, the user may choose the correct word from the list[,], or may choose to edit a similar word if the correct word is not on the list." *Id.*, 4:19-20, 16:22-24, 23:32-35. Alternatively, *Roberts* explains that a user may type or speak the first letters of the word. *Id.*, Abstract. Its recognition algorithm will then call words satisfying those letters, and the "choices displayed again" to the user. *Id.* This list may be "displayed from a large backup vocabulary," with the "best words" selected by "a statistical language model and optionally word models derived from a phonemic dictionary."⁵ *Id.* Once the correct word is identified by the user, a "speech sample is used to create or update an acoustic model for the word, without further intervention by the user." *Id.* *Roberts* explains that its system will "use[] the confirmed word ... to update the language model used by the recognition system." *Id.*, 13:44-47, Fig. 8, 18:14-19-

⁵ *Roberts* explains "such language models are well known to those skilled in the art of natural language speech recognition, and this invention would work with any of them. In the preferred embodiment, the language model is a probabilistic language model based on the counts of word bigrams" *Roberts*, 13:44-60.

31, Abstract. In this way, *Roberts* explains that “[t]he system gets more and more word models and keeps improving its performance the more it is used.” *Id.*, Abstract.

C. Overview of *Verrier*

68. EP 0622724 to *Verrier* was filed in April 1994 and published in November 1994. *Verrier*, Cover. *Verrier* discloses a stylus with a microphone and a speaker:

An improved voice communication feature is disclosed for an untethered personal stylus used with a digitizing display in a pen-based computer system. **A microphone and speaker are included in an untethered stylus in conjunction with a mechanical contact and radiative pickup mechanisms**, to enable multimedia communication between the user and the pen-based computer system, through a digitizing display.

Id., Abstract (emphasis added). I understand that the ’405 patent’s specification admits to the teaching of *Verrier* of a microphone in a stylus as prior art:

IBM’s European patent application EP 622724, published in 1994, discloses a microphone in a stylus for providing speech data to a pen-based computer system, which can be recognized by suitable voice recognition programs, to produce operational data and control information to the application programs running in the computer system.

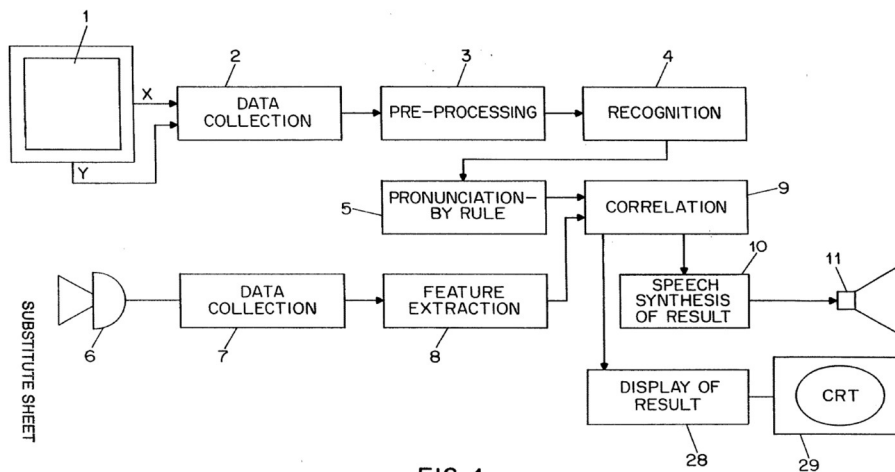
See, for example, column 18, lines 35-40 of EP 622724.

Speech data output from the microphone and received at the disclosed pen-based computer system can be converted into operational data and control information by a suitable voice recognition program in the system.

'405 patent, 2:44-54.

D. Overview of *Roth*

69. *Roth* discloses an “Audio-Augmented Handwriting Recognition” system (*Roth*, Title), in which a handwriting input (e.g., Fig. 1 below, element 1) and speech input (e.g., Fig. 1, element 6) are used to generate text (e.g., Fig. 1, elements 9, 28). This system is shown in Figure 1:



Id., Fig. 1.

70. I understand that in one exemplary embodiment, *Roth* discloses analyzing handwritten input (element 1) to derive a number of candidate words and receiving speech input (element 6) to correlate with that handwritten input to

narrow candidate words to one result before displaying the match word on the display (element 28). *Id.*, 4:12-28, 6:23-7:4, 8:11-19, 11:3-15. *Roth* also discloses the “opposite sequence” of analyzing speech input to derive a number of candidate words and receiving handwritten input to correlate with the speech input to narrow candidates to one result before displaying the result. *Id.*, 11:16-30. According to *Roth*, this system “could, in fact, apply both approaches ... and choose the result with the higher degree of confidence.” *Id.*, 11:30-32.

71. *Roth* discloses a “novel step of correlation” in which extracted features of the speech input are correlated with extracted features of the handwritten information to generate a matched candidate word. *Id.*, 6:32-7:17. *Roth* also discloses offering the user a chance to correct the candidate word either by touchpad input or speech. *Id.*, 8:15-19 (“[I]n the (hopefully rare) case that the match was incorrect, the user will so indicate, either by touchpad input or by speech input to the system.”).

IX. Claims 1-12 Are Unpatentable

A. Claims 1-3, 6-9, and 12 Would Have Been Obvious over *Waibel* (incorporating *Suhm*, *Vo*, and *Manke*) and *Roberts* (Ground 1) or over *Waibel* in view of *Suhm*, *Vo*, *Manke*, and *Roberts* (Ground 2).

72. In my opinion, *Waibel*’s multimodal recognition method (incorporating *Suhm*, *Vo*, and *Manke*), in view of *Roberts*, renders obvious each of

the claimed functions of claims 1-3, 6-9, and 12 using the same or equivalent structures and acts as the '405 patent.

73. I also believe that it would have been obvious to combine these additional references such that *Waibel* in combination with *Suhm*, *Vo*, *Manke*, and *Roberts* renders each of claims 1-3, 6-9, and 12 obvious for the same reasons.

1. Independent Claim 1

a. [1P] “*A system for generating text responsive to message input of a first type and a second type, the system comprising:*”

74. In my opinion, *Waibel* (with *Suhm*, *Vo*, and *Manke*) teaches this limitation. For example, *Waibel* discloses a system of generating text. *Waibel*, 169-70. It explains its system comprises a number of multimodal processors, including the speech recognition system from *Suhm* and the handwriting recognition system from *Manke*. *Id.*, 167. These recognizers receive different inputs, including message input of a first type (“speech” input) at the speech recognizer, and message input of a second type (“handwriting” input) at the handwriting recognizer. *Id.* These recognizers output information to the joint interpreter, which in turn “produc[es] an interpretation of user intent.” *Id.*, 167-168. This interpretation can include producing an “HTML report” or QuickDoc, which a POSITA would understand includes text. *Id.*, 169-70 (noting “audio file is also run through the speech recognizer to produce an automatic transcription”); If the

graphical user interface is available, then the user can highlight erroneous words provided by the “recognition hypothesis displayed” and “choos[e] among different correction methods.” *Id.*, 168. *See also Vo*, Figure 2, showing the parse using the output of the speech recognizer as a text string, where “[t]he text string output from the speech recognizer is processed by the Phoenix semantic parser developed by Ward” (*Vo* 3456).

- b. [1a] “(a) *a unified message model including:*
 (1) *a shared language model;*
 (2) *a first model specific to a first type of message recognition; and*
 (3) *a second model specific to a second type of message recognition;*”

75. In my opinion, *Waibel* (with *Suhm*, *Vo*, and *Manke*) renders obvious this limitation.

76. I believe that a POSITA would have found it obvious that *Waibel* teaches “*a unified message model*” that includes (1) a shared language model used in the speech recognizer, handwriting recognizer, and/or joint recognizer (*Waibel*, 167-68), (2) a model used in the recognition front-end of the JANUS speech translation system (*id.*, 167) specific to speech input, and (3) a model used in the MS-TDNN-based handwriting recognizer specific to handwriting input (*id.*), as I explain in detail below. *Waibel* explains that its technique “[r]ecogni[zes] output from the modality processors ... that are merged together to produce the combined interpretation,” which “leads to uniform handling of high-level information from

all input sources.” *Waibel*, 168. The ’405 specification explains an exemplary “unified message model” is in Figure 3B. ’405 patent, 22:52-64.

i. “(1) a shared language model”

77. The ’405 patent explains a “shared language model” is “any language model that may be shared between two types of subordinate message recognizers included within a message recognizer.” ’405 patent, 22:38-51; *see, e.g., id.*, Fig. 3B. I believe that *Waibel* (with *Suhm, Vo, Manke*) discloses or at least renders obvious such “*shared language model*.” Not only was the sharing of a language model between different recognizers well-known in the art at the time (*see, e.g., Roberts*, Fig. 1 (103, 111, 104, 180); *Starner*, Abstract; ’000 Patent, 6:27-46; *Moran* at 296), *Waibel*, in particular, was designed to favor the use of shared language models, as explained in detail below. *Id.*

78. First, *Waibel* discloses a language model used in its speech and/or handwriting recognizers. *Waibel*, 167 (speech recognizer “can be adapted to any task domain by retraining *the language models* and possibly tuning the acoustic models” (emphasis added)); *Suhm*, 222 (“standard word bigram language model”); *Manke*, 1098, 1099 (dictionaries). A POSITA would have understood the language model in *Waibel*’s speech and/or handwriting recognizers is shared or would at least have found the use of a shared language model obvious.

79. For example, *Manke* describes using handwriting language models that “were originally defined for speech recognition” and chosen to “take advantage of the *synergy effects* between handwriting recognition and speech recognition, since in our case the final goal is to integrate our speech recognizer JANUS ... and the proposed NPen⁺⁺ system into a multi-modal system.” *Manke*, 1099 (emphasis added). A POSITA would therefore have understood that the language model in the handwriting recognizer would be shared, or at least would have been obvious to be shared with the speech recognizer, and that using such a shared language model would have been *Waibel*’s preferred design to integrate the recognizers into a single multi-modal system with synergistic effects. A POSITA would have also understood that using a shared language model increases the performance and accuracy of each recognizer while simplifying computation and minimizing memory requirements.

80. Similarly, Table 1 in *Manke* shows a handwriting recognition algorithm based on a dictionary that can be changed without rewriting the recognizer. *See Manke*, 1098 (Table 1). A POSITA would have understood that a dictionary used in *Waibel*’s recognizer is a language model because of its role in limiting recognition to words a user can speak, independent of *Waibel*’s handwriting algorithm. Because this is independent of the handwriting algorithm, a POSITA would have understood that it is primed for being shared between speech

and handwriting recognizers. Thus, a POSITA would have at least found it obvious to share the dictionary language model used in the handwriting recognizer with the speech recognizer in *Waibel*, especially because the speech recognizer also needs to use a dictionary language model (*Suhm*, 223-24), and that model could be the same as that in the handwriting recognizer for simplicity and minimized memory requirements. To that end, a POSITA would have also understood that *Waibel*'s speech and handwriting recognizers were designed to accommodate the same dictionary language model, as reflected by the same way they process respective inputs, for example, by conversion into intermediate forms (phonemes and stroke units, respectively) and then into letters, as modelled by a dictionary. In *Waibel*'s system, where multiple modalities can use the same language model, it would have been unnecessary to build separate language models for speech and handwriting input.

81. This is also supported by other disclosures in *Waibel*. For example, a POSITA would have found it beneficial to share the bigram language model in the JANUS speech recognizer between the handwriting and speech recognizer. *Suhm*, 222. Even though the bigram language model is not expressly described as being used with the handwriting recognizer, a POSITA would have understood this would nevertheless be the case because the bigram statistical language models in the speech recognizer and the Viterbi search of the Hidden Markov Models in the

handwriting recognizer are compatible. *See Manke*, 1096-98; *see generally Bailey*, 1:31-35, 5:62-6:17; 8:2-36; 10:45-58; 11:9-37. This too aligns with *Waibel's* disclosed desire to achieve synergy in speech and handwriting for vocabulary management. *See, e.g., id.*, 1099. At the very least, the POSITA would have found it obvious to use a shared bigram language model to ease computation and memory requirements.

82. In fact, I believe that *Waibel's* speech recognizer was designed to favor the use of shared language models. Specifically, the speech recognizer is language-independent (*see, e.g., Suhm*, 221-22 (“Each module is language-independent in the sense that it consists of a general processor that applies independently specified knowledge about different languages.”)) and therefore model-independent such that its language model can be changed without the need to rebuild the recognizer algorithm. *See also Suhm*, 221-22.

83. Because each model is built from training data, data processing times required to build a model can become large. *See, e.g., Waibel* Table 1 (showing close to 2000 dialogs and 500,000 words for the English Language scheduling model data). Based on my experience, this is true even if the same training data can produce acoustic and language models. Performing substantially the same data processing to build multiple models for different recognizers in a system unnecessarily consumes processing times and resources. A POSITA would have

been motivated to save resources by sharing a model between speech and handwriting recognizers when possible.

84. Moreover, the speech recognizer is designed to use language models that can be changed dynamically as needed (*see, e.g., Suhm, 222*, teaching that the language model can be changed based on the discourse state) such that it may require access to more than one language model. Therefore, a POSITA would have understood that the speech recognizer is designed to be able to access multiple language models and change language models as needed such that the memory requirements for the language models may become large. Thus, I believe that a POSITA would have found it obvious to share language models whenever possible in order to reduce memory requirements.

85. In my opinion, an aspect of a shared language model is also taught in *Vo*, i.e., *Waibel's* incorporated reference [6]. *Waibel*, 168, 170. For example, *Vo* describes a dialog processor that maintains a single frame per dialog to hold information obtained from speech and handwriting inputs such that new information from each additional input event is merged into this frame. Specifically,

[a] dialog between a user and the system can span several related multimodal events, where each event may contain one or more spoken utterances and/or pen gestures. The dialog processor maintains a single frame per dialog to

hold all information obtained since the start of the dialog;
new information from each additional input event is
merged into this frame and the scores are weighted
appropriately so that frames produced by different input
events in the same dialog contribute equally to the end
result.

The domain-independent dialog processor also consults
the domain interface (i.e., the calendar in this case) to
filter out hypotheses that would cause domain errors such
as missing or conflicting parameters. If it fails to find an
acceptable hypothesis with a high enough score, it uses
the error information returned by the domain interface to
construct an appropriate feedback message to guide the
user (only this process depends on the task domain). As
soon as a good hypothesis is found, the dialog processor
sends it to the domain interface as the operation to carry
out and starts a new dialog. The user has the option of
undoing the operation just performed if it is incorrect.

Vo, 3547; *see also id.*, Figs. 1 and 2.

86. In my opinion, a POSITA would have understood or at least found
obvious that, in the context of *Waibel's* system, the particular dialog frames
described in *Vo* and used in *Waibel* that process outputs from the speech and
handwriting recognizers in *Waibel* for further recognition is indicative of another
aspect of a shared language model. Specifically, frames were a well-known natural

language processing construct reflecting both syntax and semantics. I used dialog frames in my own work as early as 1979. A frame is used to describe the intent of an utterance by having slots with different roles that can be filled during input or repair, for example, a scheduling application might use frames with slots such as date, time, location, duration, other parties, etc. Different parts of speech (syntax) fill various slots, but language understanding of the entire utterance applies semantic information, such as the difference between the time a meeting will last and the start time of the meeting. As *Vo* describes above, the frames are filled as part of a dialog in which input is requested, provided, recognized, and incorporated until a frame is complete and the system performs the intended action. This dialog directly interacts with the lower-level recognizers, i.e., both the speech and handwriting recognizers, by temporarily changing their operative language model. *See, e.g., Waibel*, 168 (“In form filling tasks, knowledge of the current field can provide powerful constraints for recognition by restricting [the] vocabulary and language model accordingly.”); *Suhm*, 222 (“After parsing, a discourse processor can be used to put the current utterance in the context of previous utterances, opening possibilities to integrate the speech and natural language processing components of the system to resolve parsing ambiguities and dynamically adapt the vocabulary and language model of the recognizer based on the current discourse state.”). A POSITA would have understood the dialog frames used in

Waibel's system to constitute a type of shared language model because they embody a dynamic language model that changes the operation of the recognizers depending on the dialog state and using shared inputs from both the speech and handwriting recognizers. *See also Suhm*, Fig. 2, Section 3.2 at 223.

87. In my opinion, even if modifications were needed to use a shared language model between the speech and handwriting recognizers, a POSITA would have found any such modifications straightforward with no technical challenges due to the similarities in the recognition techniques used in the recognizers. For example, a POSITA would have understood *Waibel*'s speech recognizer to be of the common type used at the time that decomposes words into phonemes, the smallest unit of the speech input, before building a three-state Hidden Markov Model to represent each phoneme. *See, e.g., Suhm*, 222; *see generally Bailey*, 1:31-35, 5:62-6:17; 8:2-36; 11:9-37. Similarly, the handwriting recognizer of *Waibel* uses characters, the smallest unit of handwriting input, before building a tri-state Hidden Markov Model to represent each character. *See, e.g., Manke*, 1096; *see generally Bailey*, 1:31-35, 5:62-6:17; 8:2-36; 10:45-58; 11:9-37. Accordingly, I believe that a POSITA would have recognized that the recognizers process words in the dictionary using the same type of algorithm, making the use of a shared language model nearly seamless.

ii. “(2) a first model specific to a first type of message recognition”

88. *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses this element. *Waibel* discloses an acoustic model for speech recognition. *Waibel*, 167. Indeed, its speech recognizer can be adapted “by retraining the language models and possibly tuning the *acoustic models* if the domain involves special vocabulary.” *Id.* (emphasis added). A POSITA would have also understood *Suhm* describes at least an acoustic model in the speech recognizer for recognizing speech, using for example “TDNN, MS-TDNN [3], MLP, LVQ [4], and HMM’s [5, 12] for acoustic modeling.” *Suhm*, 222.

iii. “(3) a second model specific to a second type of message recognition”

89. I believe that *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses this element. For example, a POSITA would have recognized that *Waibel*’s incorporated reference [7], *Manke*, discloses at least the following handwriting-specific models:

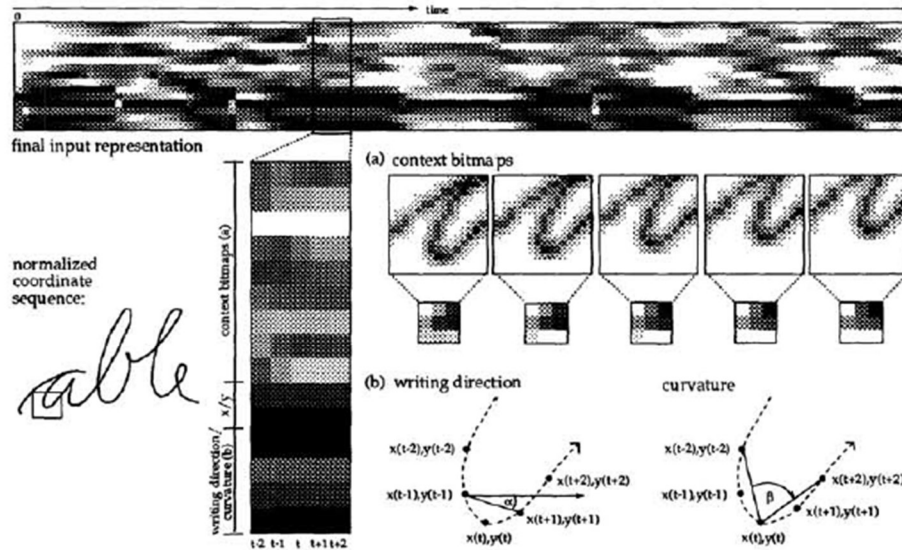


Figure 1: Feature extraction for the normalized word “able”. The final input representation is derived by calculating a 15-dimensional feature vector for each data point, which consists of a context bitmap (a) and information about the curvature and writing direction (b).

Manke, 1095 (Fig. 1) (see, e.g., “15-dimensional feature vector”).

3.1 MODELING ASSUMPTIONS

Let $W = \{w_1, \dots, w_K\}$ be a vocabulary consisting of K words. Each of these words w_i is represented as a sequence of characters $w_i \equiv c_{i_1} c_{i_2} \dots c_{i_k}$ where each character c_j itself is modelled by a three state hidden markov model $c_j \equiv q_j^0 q_j^1 q_j^2$. The idea of using three states per character is to model explicitly the initial, middle and final section of the characters. Thus, w_i is modelled by a sequence of states $w_i \equiv q_{i_0} q_{i_1} \dots q_{i_{3k}}$. In these word HMMs the self-loop probabilities $p(q_{i_j} | q_{i_j})$ and the transition probabilities $p(q_{i_j} | q_{i_{j-1}})$ are both defined to be $\frac{1}{2}$ while all other transition probabilities are set to zero.

Id., 1096 (see, e.g., three-state Hidden Markov Model).

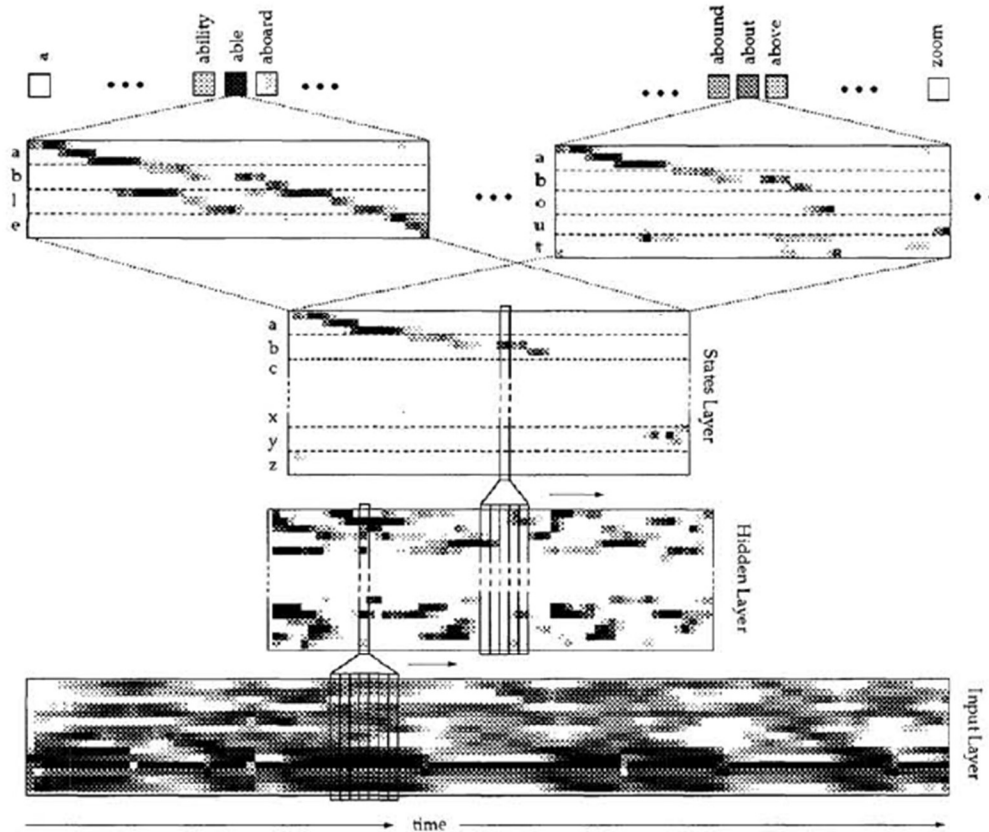


Figure 2: The Multi-State TDNN architecture, consisting of a 3-layer TDNN to estimate the a posteriori probabilities of the character states combined with word units, whose scores are derived from the word models by a Viterbi approximation of the likelihoods $p(\mathbf{x}_0^T | w_i)$.

Id., 1097 (Fig. 2) (*see, e.g.*, “Viterbi approximation” and “scores are derived from the word models”).

- c. [1b] “(b) a first message recognizer, responsive to the first type of message input to provide text data in accordance with both the shared language model and the first model; and”

90. In my opinion, *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses or at least renders obvious this limitation.

91. *Waibel* discloses a first message recognizer (the front-end of JANUS speech translation system and joint multimodal interpreter (*Waibel*, 167)) that, responsive to a first type of message input (speech input), provides text data (e.g., text output in QuickDoc (*id.*, 169-70)) in accordance with a shared language model (e.g., language model in multimodal interpreter, see discussion in limitation [1a] above) and the first model (e.g., the acoustic model in the recognition front-end of the JANUS speech translation system, see limitation [1a]). *See also Vo*, 3456, Fig. 2 (showing output of speech recognizer used by the parser is a text string and the “text string output from the speech recognizer is processed by the Phoenix semantic parser”); *supra* Section IX.A.1.b.

- d. [1c] “(c) a second message recognizer, responsive to the second type of message input to provide text data in accordance with both the shared language model and the second model”

92. In my opinion, *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses or at least renders obvious this limitation.

93. For example, *Waibel* teaches a second message recognizer (e.g., the MS-TDNN-based handwriting recognizer and the joint multimodal interpreter (*Waibel*, 167)) that, responsive to the second type of message input (e.g., handwriting input), provides text data (e.g., text in QuickDoc (*id.*, 169-70)) in accordance with both the shared language model (e.g., the language model used in the multimodal interpreter, see discussion in limitation [1a] above) and the second

model (e.g., the handwriting model used in the MS-TDNN-based handwriting recognizer, see [1a] above). Further, although the above-cited reference to *Vo*, Figure 2, specifically mentions speech recognition input, a POSITA would recognize that it would apply to handwriting input as well, because, as stated, “[a] person using our Jeanie multimodal calendar can employ any combination of spoken input, gesturing with a pen on a touch-sensitive screen, or handwritten words to interact with the system” (*Vo*, 3545), and the parser shown in Figure 2 would have been understood to be part of the “Multimodal Interpreter” which receives inputs from multiple recognizers, in Figure 1; supra Section IX.A.1.b. This is supported by *Waibel*’s teaching that “[r]ecognition output from the modality processors are parsed into partially filled frames that are merged together to produce the combined interpretation as described in [6]. This technique leads to uniform handling of high-level information from all input sources, which is very important for modularity and extensibility.” *Waibel* 168.

- e. [1d] “(d) *the shared language model is trainable responsive to user correction of misrecognition by either of the first and second message recognizers, thereby improving accuracy of each of the first and second message recognizers*”

94. In my opinion, *Waibel* (with *Suhm*, *Vo*, and *Manke*) and *Roberts* renders obvious this limitation.

95. As I opined for limitation [1a], *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses or at least renders obvious a shared language model. To the extent that *Waibel* (with *Suhm*, *Vo*, and *Manke*) does not expressly disclose that this shared model is trainable, it would have been obvious in view of *Roberts*. *Id.*

96. *Roberts* discloses a method for speech recognition and training that uses a language model that is trainable responsive to user correction of misrecognition by either of a first (e.g., speech) or second message (e.g., keystroke input) recognizer. For example, *Roberts* discloses that, after a user dictates for speech recognition, a displayed word list is generated from a large backup vocabulary “using a statistical language model and optionally word models derived from a phonemic dictionary.” *Roberts*, Abstract. When the user chooses the correct word from the word list by either “keyboard or voice command,” the language model is updated. *Id.* (“[w]hen the correct word is chosen by the user, the speech sample is used to create or update an acoustic model for the word, without further intervention by the user. As the system is used, it also constantly **updates its statistical language model**. The system gets more and more word models and keeps improving its performance the more it is used” (emphasis added)), 12:13-14:54 (describing error correction), 16:11-17 (“[t]his method only saves tokens for utterances when the user specifically confirmed the word to be associated with that utterance by means of a pick-choice command. This lets the user greatly reduce the

chance that an **incorrectly recognized utterance** will be used to train models of the wrong words.” (emphasis added)), 16:22-28 (“[i]t is true that a user, by **pressing** ‘fl’, can use a pick-choice command to pick the top choice in an utterance’s first recognition (that performed in step 123). But normally the user will not, since it is much easier to confirm such an initial top choice merely by **speaking** another text word, which will cause the branch of programming under step 108 to be performed.” (emphases added)); *see generally id.*, 15:58-16:28 (updating of acoustic model), 23:59-25:31 (updating of language model). I note that *Roberts*’s methods summarized here are specifically for correction, in addition to training, and not necessarily invoked when recognition is successful. *See, e.g., Roberts* 27:30-33 (“when any of these words occurs again later it can probably be entered with no keystrokes at all (because a correct recognition can be confirmed by beginning the next utterance...)”).

97. As illustrated in the exemplary process of Figure 1 below, *Roberts* discloses receiving speech and keystroke input (red boxes 111 and 104) and using a shared language model in speech recognition (e.g., red box 123). *Roberts* also discloses training the shared language model responsive to user correction of misrecognition (e.g., annotated blue box below) by the speech recognizer (e.g., red boxes 214, 216, 180):

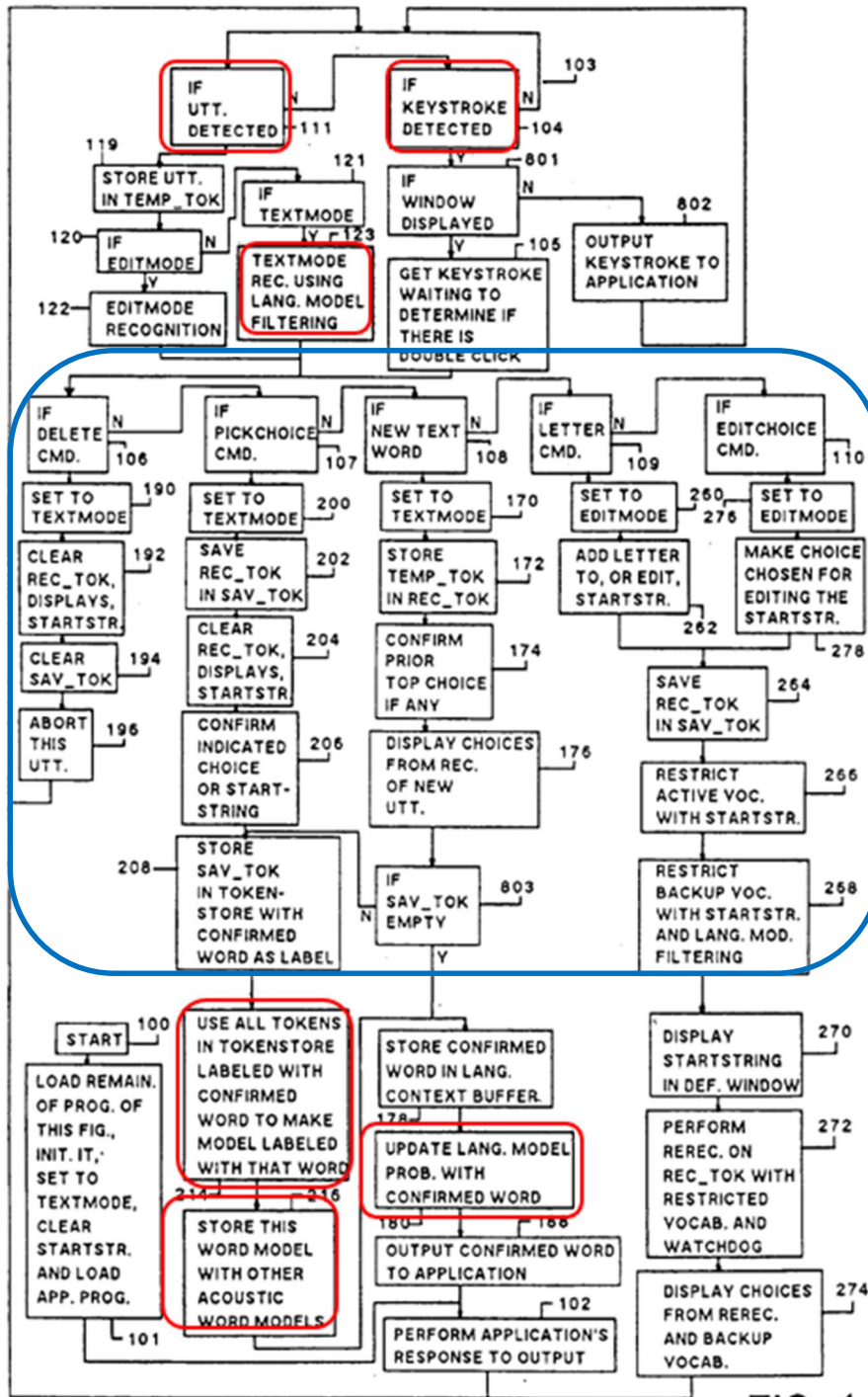


FIG. 1

Id., Fig. 1 (annotated); *see also id.*, 12:13-14:54 (describing error correction); 13:44-50 (“[a]fter step 178 stores the confirmed word, if any, in the language context buffer, step 180 uses the confirmed word, if any, to update the language

model used by the recognition system. Several such language models are well known to those skilled in the art of natural language speech recognition, and this invention would work with any of them.”), 19:23-31 (“[w]hen recognition is being performed against phonemic as well as against individually trained acoustic models, it is desirable that a language model be used to limit the recognition vocabulary to a number of words which the recognizer can handle in a response time which the user finds satisfactory. This is done by limiting the recognition vocabulary to such a number of words which the language models considers [sic] most likely to occur in the current language context.”), 7:55-56 (“[i]f an utterance is detected, step 118 forms a token of it.”), 16:7-22 (“[t]he saving of tokens associated with confirmed words and the use of multiple tokens saved in conjunction with the confirmed word performed by the program of FIG. 1 provide an effective and efficient means of adaptive speech recognition. This method only saves tokens for utterances when the user specifically confirmed the word to be associated with that utterance by means of a pick-choice command. This lets the user greatly reduce the chance that an incorrectly recognized utterance will be used to train models of the wrong words. It also tends to greatly reduce unnecessary training, since normally the pick-choice command is only used to pick words other than the top-choice, and thus it reduces the likelihood that the system will replace word models that are producing correct recognition.”), 29:55-30:53 (“[w]hen the

user speaks a multiword phrase [and] the system fails to correctly recognize it, he has two major ways to get the correct phrase entered with the embodiment shown in FIGS. 25 and 26. The first is to enter all the characters of the phrase with letter commands, either by typing them or by speaking them with the communications alphabet The second method of correcting an incorrectly recognized multiword phrase is to use the phrase edit routine of FIG. 26.”), claim 48 (“[a]n adaptive speech recognition method for recognizing each of a succession of spoken words over a period of time and for improving a set of acoustic models used during that recognition”).

98. In my opinion, a POSITA would have been motivated to incorporate *Roberts*’s method of training a language model responsive to user correction of misrecognition in the multimodal system of *Waibel* so that *Waibel* can similarly train its shared language model responsive to user correction of misrecognition in the speech or handwriting recognizer. As explained in *Roberts*, this method allows a user to “begin productive work on large vocabulary, natural language dictation tasks with little or no advance training of word models” (*id.*, 2:8-12), “build acoustic models from tokens that are collected during actual use of the system rather than during a separate training or enrollment process” (*id.*, 2:17-22), “easily correct errors made in continuous speech recognition” (*id.*, 2:23-25), and “develop a language model for use in a speech recognition system which is customized to

his or her particular dictation, without requiring much additional effort to obtain such customization” (*id.*, 2:26-30). Accordingly, a POSITA would have appreciated and expected the same benefits in *Waibel*’s system.

99. *Roberts* also discloses that the training of language models improves the accuracy of the speech recognizer. *See, e.g., id.*, 2:5-7 (“[i]t is an object of the present invention to provide a speech recognition system which improves the speed and accuracy with which users can dictate desired text.”), Abstract (“[w]hen the correct word is chosen by the user, the speech sample is used to create or update an acoustic model for the word, without further intervention by the user. As the system is used, it also constantly updates its statistical language model. The system gets more and more word models and keeps improving its performance the more it is used.”), 2:17-22 (“[i]t is a further object of this invention to provide improved speech recognition performance by providing an improved means to build acoustic models from tokens that are collected during actual use of the system rather than during a separate training or enrollment process.”). Accordingly, I believe that a POSITA would have found it obvious that the combination of *Roberts* and *Waibel* would also improve the accuracy of the speech recognizer in *Waibel*.

100. Moreover, at least because the handwriting recognizer in *Waibel* includes the language model that would be trained and improved in view of

Roberts, in my opinion, the improved language model necessarily leads to a more accurate handwriting recognizer.

101. Furthermore, I believe that the multi-faceted compatibilities of *Waibel* and *Roberts* would have made any combination well within the grasp of a POSITA yielding entirely predictable results and requiring no undue experimentation. For example, *Waibel* discloses a standard word bigram model (*Suhm*, 222), which is a compatible format with *Roberts* (see *Roberts*, 13:44-60). Further, *Waibel* discloses a similar method of error correction to that discussed in *Roberts*, including using a GUI to highlight and correct errors in speech recognition. See, e.g., *Waibel*, 168 (“[E]rrors have to be identified, either initiated by the system, for instance based on some confidence measure, or by the user. If a graphical user interface (GUI) is available, the user can simply highlight erroneous words (the reparandum) in the recognition hypothesis displayed. For error correction, the user provides additional input, choosing among different correction methods: repair by repeating the reparandum by respeaking, spelling out loud, or handwriting; repair by paraphrasing the reparandum; repair by pen gestures (e.g. to delete or insert), in addition to the standard repair by typing or selecting among N-best alternatives [R]equiring the user to identify errors by highlighting them is appropriate, since the user naturally focuses on the actual recognition hypothesis.”); *Roberts*, 4:15-20 (“The system also enables the user to select an individual word from a given

indicated best matching word sequence, and to correct that word sequence by correcting the selected word. Preferably the word to be corrected can be selected by voice.”). Furthermore, I believe that *Roberts*’s method of improving Hidden Markov Models and language models with training data also applies to *Waibel*’s recognizers because *Waibel*’s speech and handwriting recognizers both use algorithms based on Hidden Markov Models (e.g., JANUS speech translation system and NPen⁺⁺) that can be improved and user-adapted through additional training data. *See, e.g., Waibel*, 167 (“[w]e show how cross-modal cues can be used effectively to recover from errors or miscommunications, since confusions in one modality usually are unambiguous in another”), 168 (“[t]he potential for error with any speech interface will be balanced by the redundant and alternate ability to enter information or queries by spelled, handwritten or typed input. Rather than fighting a particular recognition error[,] the user can therefore quickly circumvent the problem by choosing a different modality. Such alternates not only eliminate user frustration with potentially recurring errors, but also provide an avenue for background learning and adapting.”); “Given a set of error recovery methods feasible with current technology in a given application context, a crucial issue is to predict which methods a user will prefer.”); *Roberts*, 2:23-25 (“[i]t is still another object of this invention to provide a system which enables a user to easily correct errors made in continuous speech recognition.”); *see generally Bailey*, 8:2-36.

Thus, in my opinion, a POSITA would have had no trouble combining *Waibel* and *Roberts*.

2. Claim 2

“The system of claim 1 wherein:

(a) the first type of message recognition is speech recognition, the first type of message input is voice input, and the first model is an acoustic model; and

(b) the second type of message recognition is handwriting recognition, the second type of message input is freehand input, and the second model is a handwriting model.”

102. In my opinion, *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses these additional limitations. As I discussed above for limitations 1[a] and 1[b], *Waibel* discloses its first type of message recognition is speech recognition, with the first type of message input being voice input, and the first model being an acoustic model. *E.g.*, *Waibel*, 167 (“Speech: Our speech recognition subsystem is based on the recognition front-end of the JANUS speech translation system [5] which is capable of processing speaker-independent, spontaneous speech. The recognizer can be adapted to any task domain by retraining the language models and possibly tuning the acoustic models if the domain involves special vocabulary.”); *see also* discussion above for elements [1a] and [1b]; *see supra* Sections IX.A.1.b, IX.A.1.c.

103. Furthermore, as I discussed above for limitations 1[a] and 1[b], *Waibel* likewise discloses that the second type of message recognition is handwriting recognition, with the second type of message input being handwriting input. *See, e.g., Waibel*, 167 (“[o]ur MS-TDNN-based handwriting recognizer [7] is capable of processing writer-independent, continuous (cursive) handwriting at a recognition rate of 94% on a 20,000-word vocabulary.”). Furthermore, a POSITA would have understood that the model used in the MS-TDNN-based handwriting recognizer is a handwriting model. *See* discussion above for elements [1a] and [1c].

3. Claim 3

“The system of claim 2 further comprising a stylus, wherein the shared language model is trainable responsive to user correction of misrecognition by the first message recognizer by (a) movement of the stylus across visual indicia of dictated text that has been erroneously generated to specify a selection from the text, and (b) correction of the selection.”

104. *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses using a stylus (e.g., pen) to correct misrecognition by the recognizer, which I believe that a POSITA would have understood to include misrecognition by the speech recognizer. *See, e.g., Waibel*, 168 (“[R]equiring the user to identify errors by highlighting them is appropriate, since the user naturally focuses on the actual recognition hypothesis. For error correction, we offer to repeat the input (potentially switching to another modality), and pen gestures to indicate where to insert words, or which word to

delete.”; “If a graphical user interface (GUI) is available, the user can simply highlight erroneous words (the reparandum) in the recognition hypothesis displayed. For error correction, the user provides additional input, choosing among different correction methods: repair by repeating the reparandum by respeaking, spelling out loud, or handwriting; repair by paraphrasing the reparandum; repair by pen gestures (e.g. to delete or insert), in addition to the standard repair by typing or selecting among N-best alternatives.”).

105. In my opinion, a POSITA would also have understood that *Waibel*’s correction of misrecognitions through, e.g., highlighting of erroneous words and “repair by pen gestures (e.g. to delete or insert)” (*Waibel*, 168) constitutes user correction of misrecognition by the speech recognizer by (a) movement of the stylus across visual indicia of dictated text that has been erroneously generated to specify a selection from the text, and (b) correction of the selection.

106. And as I explained above, *Roberts* teaches a system with a shared language model trainable responsive to user correction of misrecognition by the first message recognizer (e.g., speech recognizer) by either a user’s speech or keyboard input correction of a selection. *Supra* Section IX.A.1.e. Thus, I believe that a POSITA would have found it obvious to train the shared model in the system of *Waibel* modified and/or incorporated in view of *Roberts* (see limitation [1d], *supra* Section IX.A.1.e) responsive to user correction of misrecognition in the

manner described in *Waibel*, e.g., using a stylus. *See, e.g., Roberts*, 4:15-20 (“The system also enables the user to select an individual word from a given indicated best matching word sequence, and to correct that word sequence by correcting the selected word. Preferably the word to be corrected can be selected by voice.”), Abstract (“When the correct word is chosen by the user, the speech sample is used to create or update an acoustic model for the word, without further intervention by the user. As the system is used, it also constantly updates its statistical language model. The system gets more and more word models and keeps improving its performance the more it is used.”).

107. In my opinion, this would have been especially obvious because *Roberts* also teaches that the correction in *Roberts* can be performed using different modalities, such as speech or a keyboard. *Id.*, 2:48-55 (“Preferably the menu of character sequences is a list of best matching words produced by speech recognition of one of the training utterances, and the user can select and edit the desired menu choice by voice. It is also preferred that the system is used as part of a speech recognition system which produces as its output for each spoken word either the word recognized or the word edited and selected for that spoken word.”), Abstract (“By keyboard or voice command, the user may choose the correct word from the list or may choose to edit a similar word if the correct word is not on the list.”).

4. Claim 6

“The system of claim 1 wherein the shared language model comprises a syntactic model and a semantic model.”

108. As discussed in connection with element [1a] above (*supra* Section IX.A.1.b), I believe that *Waibel* (with *Suhm*, *Vo*, and *Manke*) teaches a shared language model in its joint interpretation system as described in *Vo. Waibel*, 168. I also believe that a POSITA would have understood that the interpretation method and system used in *Waibel* and described in *Vo* comprises a syntactic model and a semantic model. First, *Waibel* discloses a bigram language model, which is a syntactic model because it describes probabilities of words appearing in sequence, which is syntactic information.⁶ Second, a semantic model is disclosed in the

⁶ Per the '405 patent, a bigram model is a syntactic model ('405 Patent, 22:49-51 “A language model may be of a conventional type, for example consisting of an N-gram syntactic model”). Although *Suhm* uses a bigram model for word recognition, in the context of the JANUS system that translates full spoken sentences into another language rather than perform simple word-to-word translation, *Suhm* favors a semantic model over a specialized type of syntactic model that analyzes a full sentence into its constituent parts and considers the syntax of the target language to create an intelligible translation, a type of syntactic model commonly

dialog frames described in, e.g., *Vo*, 3546 (“The text string output from the speech recognizer is processed by the Phoenix semantic parser The parser compiles a grammar specifying semantically meaningful fragments of text into an efficient recurrent transition network that identifies these fragments in the recognized utterance. It is capable of skipping unknown words and unmatched fragments and can therefore deal gracefully with ungrammatical sentences. The concepts identified by the parser serve to determine which slots of the speech frame should be filled and what scores should be assigned to the slot values.”); *see also Suhm* at 224 (“It uses the Phoenix parser to parse input into slots in semantic frames”). As explained above in connection with limitation [1a], this shared language model receives inputs from speech and handwriting recognizers and, in response to the calculated dialog state, can change the active language models in the recognizers to improve recognition. Thus, a POSITA would have understood the frames are a

used in full-sentence translation systems (*see Suhm*, Abstract). Consistent with the ’405 patent, however, a POSITA would understand that, because syntax is the aspect of language which constrains how words can be combined in a particular order, and bigrams specifically model constraints on combinations of pairs of words, a bigram model is a basic syntactic model. *See also*, e.g., U.S. Pat. No. 5,839,106 to *Bellegarda*, 1:41-46 (cited in the ’405 Patent, 4:19-34).

semantic model as they relate to the meaning or understanding of the user's intent with the original input.

5. Independent Claim 7

- a. [7P] ***“A method for performing message recognition with a shared language model, the method comprising:”***

109. Independent claim 7 is substantially identical to claim 1, except it is directed to a method. For the reasons I discussed above for limitations [1P] and [1a], *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses or at least renders obvious this limitation. Supra Section IX.A.1.a and Section IX.A.1.b.

- b. [7a] ***“(a) performing message recognition of a first type, responsive to a first type of message input, to provide text data in accordance with both the shared language model and a first model specific to the first type of message recognition;”***

110. This limitation is substantially identical to limitations [1a] and [1b]. For the reasons I discussed for limitations [1a] and [1b], *Waibel* (with *Suhm*, *Vo*, and *Manke*) discloses or at least renders obvious this limitation. Supra Section IX.A.1.b and Section IX.A.1.c.

- c. [7b] ***“(b) performing message recognition of a second type, responsive to a second type of message input, to provide text data in accordance with both the shared language model and a second model specific to the second type of message recognition; and”***

111. This limitation is substantially identical to limitations [1a] and [1c]. For the reasons discussed for limitations [1a] and [1c], *Waibel* (with *Suhm*, *Vo*, and

Manke) discloses or at least renders obvious this limitation. Supra Section IX.A.1.b and Section IX.A.1.d.

- d. [7c] ***“(c) training the shared language model responsive to user correction of error in message recognition of either of the first and second types, thereby improving accuracy of each of the first [and] second types of message recognition.”***

112. This limitation is substantially identical to limitations [1a] and [1d].

For the reasons discussed for limitations [1a] and [1d], *Waibel* (with *Suhm*, *Vo*, and *Manke*) in view of *Roberts* renders obvious this limitation. Supra Section IX.A.1.b and Section IX.A.1.e.

6. Claim 8

“The method of claim 7 wherein:

(a) performing the first type of message recognition comprises performing speech recognition responsive to voice input in accordance with the shared language model and a first model that is an acoustic model; and

(b) performing the second type of message recognition comprises performing handwriting recognition responsive to freehand input in accordance with the shared language model and a second model that is a handwriting model.”

113. This claim is substantially identical to claim 2. For the reasons discussed above for claim 2, *Waibel* (with *Suhm*, *Vo*, and *Manke*) in view of *Roberts* renders obvious this limitation. Supra Section IX.A.2.

7. Claim 9

“The method of claim 8 wherein training the shared language model includes (a) moving a stylus across visual indicia of dictated text that has been erroneously generated to specify a selection from the text, and (b) correcting the selection.”

114. This claim is substantially identical to claim 3. For the reasons discussed above for claim 3, I believe that *Waibel* (with *Suhm*, *Vo*, and *Manke*) in view of *Roberts* renders obvious this limitation. Supra Section IX.A.3.

8. Claim 12

“The method of claim 7 wherein performing message recognition in accordance with the shared language model comprises performing message recognition in accordance with a syntactic model and a semantic model.”

115. This claim is substantially identical to claim 6. For the reasons discussed above for claim 6, I believe that *Waibel* (with *Suhm*, *Vo*, and *Manke*) in view of *Roberts* renders obvious this limitation. Supra Section IX.A.4.

B. Claims 4, 5, 10, and 11 Would Have Been Obvious over *Waibel*, *Roberts*, and *Verrier* (Ground 3) or over *Waibel*, *Suhm*, *Vo*, *Manke*, *Roberts*, and *Verrier* (Ground 4)

1. Claim 4

“The system of claim 3 wherein:

(a) the stylus includes a pointing interface and a microphone;

(b) the system further comprises a tablet responsive to the pointing interface;

(c) the voice input is from the microphone; and

(d) the freehand input is from the pointing interface being manually manipulated on the tablet.”

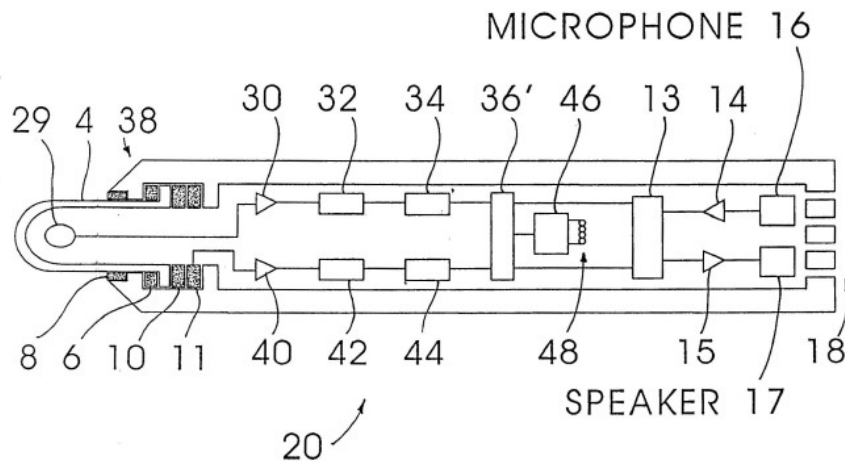
116. *Waibel* (with *Suhm*, *Vo*, *Manke*) in view of *Roberts* and *Verrier* renders these additional limitations obvious. First, I believe that *Waibel* discloses or at the very least renders obvious a “*stylus includes a pointing interface.*” *Waibel* discloses using a stylus, e.g., a pen, for handwriting input into a computer in human-computer interactions. A POSITA would have understood that such a stylus would include a pointing interface to interact with the computer receiving the input. *Waibel*, 167 (“computer agents that use speech, gesture, handwriting, pointing, spelling jointly for more robust, natural and flexible human-computer interaction”; “MS-TDNN-based handwriting recognizer [7] is capable of processing writer-independent, continuous (cursive) handwriting at a recognition rate of 94% on a 20,000-word vocabulary [and w]e employ simple heuristics to decide when to invoke handwriting recognition on pen input ...”), 168 (when “a graphical user interface (GUI) is

available, the user can simply highlight erroneous words (the reparandum) in the recognition hypothesis displayed. For error correction, the user provides additional input, [including] repair by pen gestures (e.g. to delete or insert), in addition to the standard repair by typing or selecting among N-best alternatives.”).

117. As to “*the stylus includes ... a microphone,*” and “*voice input is from the microphone,*” the limitations are not expressly disclosed in *Waibel*. They were, however, well-known stylus features at the time. They would have been obvious to a POSITA over *Waibel* (with *Suhm, Vo, Manke*) in view of *Roberts* and *Verrier*.

118. Indeed, *Verrier* discloses a stylus with a microphone for receiving voice input. *Verrier*, Abstract (noting “[a] microphone and speaker are included in an untethered stylus ... to enable multimedia communication between the user and the pen-based computer system, through a digitizing display”), 18:35-40, Fig. 2. *Verrier*’s stylus is shown below in Figure 2:

FIG. 2



119. It would have been obvious to a POSITA to incorporate a microphone in *Waibel's* pen and receive voice input at the microphone for the acquisition of high-quality speech in a pen-based system to accommodate voice communications at the user-stylus interface. *Verrier*, 3:27-35. Indeed, a POSITA would have known that placing the microphone in the stylus improves the quality of the audio signal due to the user's proximity to the stylus while dictating regardless of the location of the computer device, which in turn improves speech recognition. Even the '405 patent admits "[i]t has *long been known* to incorporate a microphone into a stylus for the acquisition of high-quality speech in a pen-based" system. '405 patent, 2:54-64 (emphasis added). A POSITA would also have known that incorporating a microphone into *Waibel's* stylus would have been a simple substitution of known elements according to their known functions to achieve predictable results. A POSITA would have been so motivated to minimize overall hardware and power

consumption while increasing system portability. The POSITA would have had a reasonable expectation of success in doing so.

120. Regarding limitation (b), the '405 patent defines a "tablet" as "any suitable hardware and/or software for performing functions of, inter alia, a speech recognizer and an editor." '405 patent, 6:11-14. As such, I believe that *Waibel* teaches a tablet, e.g., a computer agent with hardware and/or software for performing functions of a speech recognizer and an editor. *Waibel* teaches that the tablet is responsive to the pointing interface. *Waibel*, 167 ("[t]o build computer systems that operate more flexibly and robustly, and that are more intuitive to use by anyone, computer interfaces must be able to understand and process this multiplicity of human communication cues"; "[t]he need for more intuitive interfaces becomes all the more pressing as multimedia presentation, that is output, is commonly available"; "[i]n this paper we describe information agents that use multimodal interfaces to access, manipulate, create and disseminate information"; "[w]e demonstrate computer agents that use speech, gesture, handwriting, pointing, spelling jointly for more robust, natural and flexible human-computer interaction"; "we seek effective methods by which human users can easily interact with a computer by speaking, pointing, drawing, spelling"; "[o]ur MS-TDNN-based handwriting recognizer [7] is capable of processing writer-independent, continuous (cursive) handwriting at a recognition rate of 94% on a 20,000-word vocabulary.

We employ simple heuristics to *decide* when to invoke handwriting recognition on pen input”).

121. Regarding limitation (d), *Waibel* teaches using a pen for providing handwriting input on a tablet, as I explained above with respect to limitations (a) and (b) above. In my opinion, a POSITA would have understood that handwriting input implicitly includes freehand input from the pointing interface of the pen being manually manipulated on the tablet.

2. Claim 5

“The system of claim 4 wherein the visual indicia of dictated text appears on the tablet.”

122. In my opinion, *Waibel* (with *Suhm, Vo, Manke*) in view of *Roberts* and *Verrier* renders this obvious. As I explained in connection with claims 3 and 4 above, *Waibel* teaches the use of a tablet as a human-computer interface and the display of visual indicia of dictated text on the tablet. *See, e.g., Waibel*, 167, 168 (“[i]f a graphical user interface (GUI) is available, the user can simply highlight erroneous words (the reparandum) in the recognition hypothesis displayed. For error correction, the user provides additional input, choosing among different correction methods: repair by repeating the reparandum by respeaking, spelling out loud, or handwriting; repair by paraphrasing the reparandum; repair by pen gestures (e.g. to delete or insert), in addition to the standard repair by typing or selecting among N-best alternatives.”).

123. I believe that a POSITA would have understood that *Waibel* teaches using the tablet that responds to the stylus as the tablet on which visual indicia of dictated text appears. For example, *Waibel* teaches a navigation system on which information is both displayed and modified by voice, indicating a display that both responds to touch input and shows visual indicia of dictated text:

For example for navigation within visually presented image data, the user typically modifies the view by operations such as zooming and panning. **The integration of speech and gesture** increases flexibility by allowing different attributes of **an operation to be specified in whatever modality is more appropriate for each attribute**; for instance, the effect of a **spoken ‘zoom in to this <circle> area’ accompanied by a circle drawn around the desired area** is much more difficult to achieve using a single input modality.

Waibel, 169 (emphases added). Specifically, a POSITA would have understood that the display for the navigation system responds to the stylus (e.g., circle drawing) and displays dictated text (e.g., “[t]he integration of speech and gesture”; “spoken ‘zoom in to this <circle> area’”).

124. At the very least, in my opinion, a POSITA would have found this limitation obvious. Between the publication date of *Waibel* in 1997 and the ’405 patent’s filing date in 2002, the fast and commonplace development of touch-

sensitive screens in portable devices (e.g., tablets) made it obvious that the integration of various functions, including touch input and display, into a single integrated device, rather than two distinct devices, was more desirable for both user-portability, power efficiency, and manufacturing cost. A POSITA would have a reasonable expectation of successfully doing so.

3. Claim 10

“The method of claim 9 wherein:

(a) performing speech recognition responsive to voice input includes responding to voice input from a microphone in the stylus; and

(b) performing handwriting recognition responsive to freehand input includes responding to manual manipulation of the stylus on a tablet.”

125. For the reasons discussed above for claims 1, 4, and 7, I believe that *Waibel* in view of *Roberts* and *Verrier* renders obvious additional limitations of claim 10. Supra Sections IX.A.1, IX.B.1, IX.A.5.

4. Claim 11

“The method of claim 10 wherein the visual indicia of dictated text appears on the tablet.”

126. This claim is substantially identical to claim 5. For the reasons discussed above for claim 5, I believe that *Waibel* in view of *Roberts* and *Verrier* renders obvious this limitation. Supra Section IX.B.2.

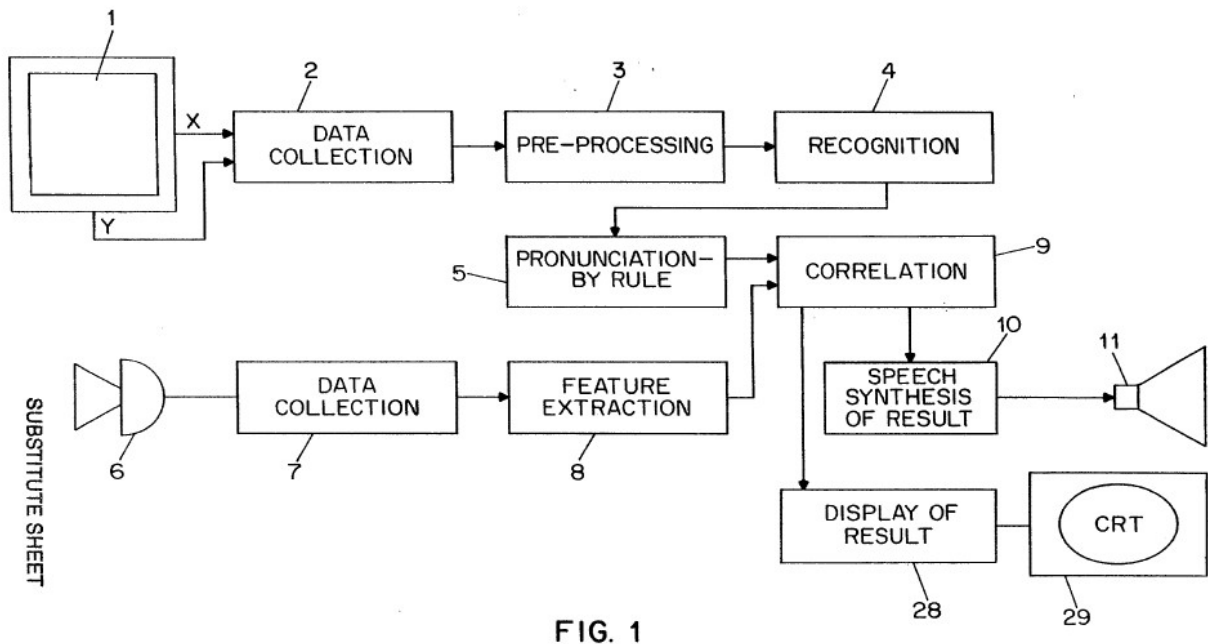
C. Claims 1-3 and 7-9 Would Have Been Obvious over *Roth* and *Roberts* (Ground 5)

127. In my opinion, *Roth* in view of *Roberts* renders obvious each of the claimed functions of these claims using the same or equivalent structures and acts as the '405 patent.

1. Independent Claim 1

- a. [1P] “*A system for generating text responsive to message input of a first type and a second type, the system comprising:*”

128. In my opinion, *Roth* teaches this limitation. For example, *Roth* discloses a system for generating text in response to message inputs. *E.g.*, *Roth*, Fig. 1, element 28 (display of result), 8:12-13 (“may be a visual display of the match character as indicated by block 28”). *Roth* explains that speech can be collected “by microphone 6 and audio storage means 7” (a “*message input of a first type*”). *Id.*, Fig. 1, element 6, 6:26-29. And handwritten input can be provided by the user at “touchpad 1” (a “*second type*” of message input). *Id.*, Fig. 1, element 1, 4:12-28. *Roth* then explains that, in response to these inputs, its system generates a “visual display of the match character” (i.e., text). *Id.*, Fig. 1, element 28, 8:3-19. These steps are shown in the block diagram provided in Figure 1:



Id., Fig. 1

- b. [1a] “(a) a unified message model including:
(1) a shared language model;
(2) a first model specific to a first type of message recognition; and
(3) a second model specific to a second type of message recognition;”

129. In my opinion, *Roth* teaches this limitation.

130. For example, *Roth* discloses a “*unified message model*” with its disclosure of element 2 (data collection), element 3 (pre-processing), element 4 (recognition), element 5 (pronunciation by rule), element 7 (speech data collection), element 8 (speech feature extraction), and element 9 (correlation), from Figure 1:

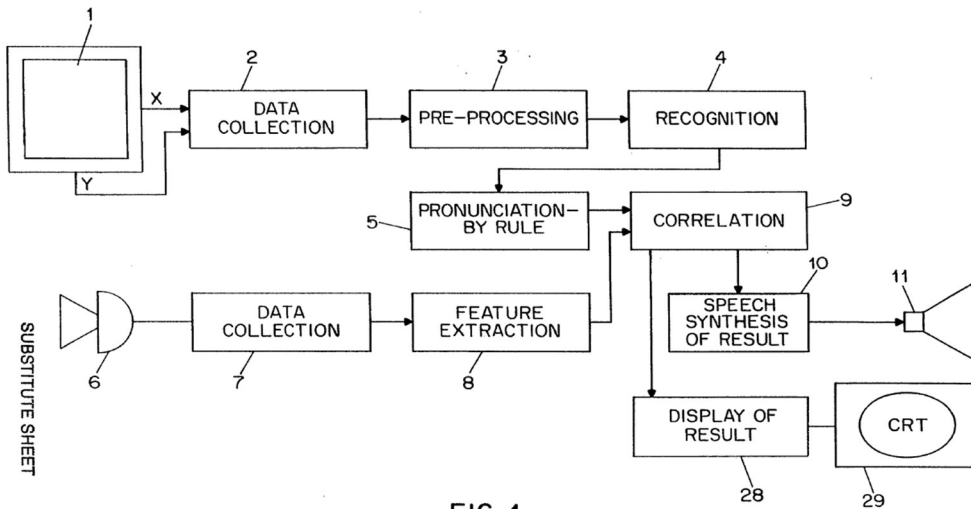


FIG. 1

Id., Fig. 1

i. (1) a shared language model

131. In my opinion, *Roth* discloses a shared language model, specifically in elements 5 and 9 from Figure 1 (the pronunciation-by-rule algorithm and correlation boxes). In elements 5 and 9, *Roth*'s system performs correlations and rankings using thresholds and margins based on features extracted from speech and handwriting input. Specifically, *Roth* discloses a language model including a pronunciation-by-rule algorithm that determines the pronunciation of a word by a rule, followed by a correlation of that word written to its pronunciation as uttered by a user. *See, e.g., Roth*, 3:22-24, 6:8-19, 10:13-25. A POSITA would have understood such a pronunciation-by-rule algorithm is a shared language model since, instead of being used only for handwriting or speech recognition, it is used as a higher-level model to measure the similarity of words spoken and written. This is performed independent of the underlying acoustic or pen stroke data and

before determining a best match from the shared input of word hypotheses independently produced by the speech and pen recognizers.

132. As another example, *Roth* teaches the use of a template store that contains certain speech templates. *See, e.g., Roth*, 9:14-24, 10:13-25. This template store and its speech templates are used for both speech and handwriting recognition, because in addition to correlations with speech inputs, the speech templates are used for correlations with word candidates obtained through the handwriting recognizer. *Id.* This template store and its correlation constitutes a shared language model because it is used to characterize what may be said or written independently from either the speech or the handwriting recognizer, and further used to correlate the outputs of the two recognizers to achieve higher accuracy.

133. Should *Roth* be found to not teach this limitation in view of limitation 1(d) (“shared language model is trainable responsive to user correction of misrecognition by either of the first and second message recognizers ...”), in my opinion, it would have been obvious in view of *Roberts*. As I explained in detail below in connection with limitation [1d], a POSITA would have found it obvious to incorporate a shared bigram language model, as taught in *Roberts*, into the system of *Roth*. *Infra* Section IX.C.1.e.

ii. (2) a first model specific to a first type of message recognition

134. In my opinion, *Roth* teaches (2) a first model specific to a first type of message recognition, e.g., speech recognition. For example, *Roth* discloses using speech models based on American and British speech templates, as well as men and women speech templates. *Roth*, 6:19-22. As another example, *Roth* also teaches the extraction of features from certain speech input, like a “determination of the amount of speech energy within various frequency ranges as a function of time.” *Id.*, 6:23-31. A POSITA would have understood the purpose of such extracted features would be for correlation with a speech model. *Id.*, 6:29-31.

iii. (3) a second model specific to a second type of message recognition

135. In my opinion, *Roth* discloses or at least renders obvious this limitation. For example, *Roth* discusses handwriting models used in existing handwriting recognizers. *E.g.*, *Roth* 1:18-29, 2:9-15. Because it does not suggest the models in its handwriting recognizer deviate from existing models, and does not otherwise provide detail on the models used in its handwriting recognizer (element 2 (data collection), element 3 (pre-processing), element 4 (recognition)), a POSITA would have understood that similar handwriting models are used in the handwriting recognizer of *Roth* to decipher handwriting. *See, e.g., id.*, 4:12-6:22. Indeed, using handwriting models in recognizers was well known at the time., as admitted in the '405 patent itself. *See, e.g.*, '405 patent, 6:22-34 (citing *Beemink*,

U.S. Patent No. 5,903,668, 10:11-32 (discussing handwriting models including “CHARPARTs)).

- c. [1b] “(b) a first message recognizer, responsive to the first type of message input to provide text data in accordance with both the shared language model and the first model; and”

136. In my opinion, *Roth* teaches this limitation.

137. For example, *Roth* teaches a first message recognizer, e.g., speech recognizer including elements 6 (microphone), 7 (data collection), 8 (feature extraction), 9 (correlation), responsive to the first type of message input, e.g., speech input via element 6, to provide text data (e.g., element 28 (display of result)), in accordance with both the shared language model, e.g., element 9 (correlation), and the first model, e.g., 7 (speech data collection) and 8 (speech feature extraction). *See, e.g., Roth*, 6:23-31; *see also id.*, 11:33-13:35, 6:33-8:2, 10:13-11:32, 14:26-17:2, 18:14-21.

138. Should *Roth* be found to not teach this limitation in view of limitation 1(d) (“shared language model is trainable responsive to user correction of misrecognition by either of the first and second message recognizers ...”), in my opinion, it would have been obvious in view of *Roberts*, as I explained in detail below in connection with limitation 1(d). *Infra* Section IX.C.1.e.

- d. [1c] “(c) a second message recognizer, responsive to the second type of message input to provide text data in

***accordance with both the shared language model and
the second model”***

139. In my opinion, *Roth* discloses or at least renders obvious this limitation.

140. For example, *Roth* discloses a second message recognizer, e.g., handwriting recognizer including elements 1 (touchpad), 2 (data collection), 3 (pre-processing), 4 (recognition), 5 (pronunciation by rule), and 9 (correlation), responsive to the second type of message input, e.g., handwriting input via element 1, to provide text data (e.g., element 28 (display of result)), in accordance with both the shared language model, e.g., element 9 (correlation), and the second model, e.g., 2 (data collection), 3 (pre-processing), 4 (recognition), 5 (pronunciation by rule). *See, e.g., Roth*, 4:29-5:5; *see also id.*, 7:5-8:2, 10:13-11:32, 14:26-17:2, 18:14-21.

141. Should *Roth* be found to not teach this limitation in view of limitation 1(d) (“shared language model is trainable responsive to user correction of misrecognition by either of the first and second message recognizers ...”), in my opinion, it would have been obvious in view of *Roberts*, as I explained in detail below in connection with limitation 1(d). *Infra* Section IX.C.1.e.

- e. [1d] ***“(d) the shared language model is trainable responsive to user correction of misrecognition by either of the first and second message recognizers,***

thereby improving accuracy of each of the first and second message recognizers”

142. Should *Roth* be found to not teach this limitation or that *Roth*’s pronunciation-by-rule/speech templates correlations are found to be not trainable shared language models responsive to user correction of misrecognition, in my opinion, this limitation would have been obvious in view of *Roberts*.

143. As I explained in connection with element [1d] under Grounds 1 and 2 above (*Waibel*), *Roberts* teaches a shared language model that is trainable responsive to user correction of misrecognition by either of a first (e.g., speech) and second message (e.g., keystroke input) recognizer. For example, *Roberts* teaches a “statistical language model and optionally word models derived from a phonemic dictionary” that generates words for matching to a user’s dictation. *Roberts*, Abstract, 7:55-57, 12:7-22, 13:44-50, 16:11-17, 16:22-28, 19:23-31, 28:61-29:32, Fig. 1. When the user chooses the correct word by either “keyboard or voice command,” the language model and acoustic model for the word are updated. *Id.* *Roberts* teaches that this training improves the accuracy of the speech recognizer. *Id.*; see *supra* Section IX.A.1.e.

144. In my opinion, a POSITA would have been motivated to use *Roberts*’s trainable language model, like “the probabilistic language model based on the counts of word bigrams” (*Roberts*, 13:50-53), in *Roth*’s language model to achieve better recognition accuracy. For example, a POSITA would have found it

beneficial to include *Roberts*'s trainable language model, such as "the probabilistic language model based on the counts of word bigrams" (*Roberts*, 13:50-53), in *Roth*'s system. Bigram models were well known in the art to effectively reduce errors in recognizers by predicting the likelihood that a second word follows a first word. *Roberts*, 13:44-60. Because this prediction is predicated on the accuracy of the first word, the bigram model would work particularly well in *Roth*'s system in which the first word is already recognized with a high degree of certainty by the pronunciation-by-rule algorithm and through a user's optional correction of a misrecognition. *Roth*, 3:22-24, 6:8-19, 8:17-19, 10:13-25. This further reduction of errors would also reduce the time it takes to provide a correct result.

145. In my opinion, a POSITA would have recognized that a bigram model in *Roth*'s system would have been especially beneficial when processing sentences. Specifically, *Roth*'s embodiments focus on processing single word inputs. A POSITA would have recognized, however, that when processing sentences, more sophisticated language models such as the bigram language model in *Roberts* would have provided additional error correction benefits because sentences inherently provide more lexical context, and bigram models use context to reduce errors. Moreover, as *Roth* already recognizes the inadequate accuracy of recognition for speech or handwriting (*see, e.g., Roth*, 1:30-2:31), a POSITA would have been motivated by *Roth*'s concern for misrecognition and would have

sought to further improve recognition accuracy by incorporating well-known and commonly used means of language models, e.g., bigram models.

146. For at least these reasons, I believe that a POSITA would have been motivated to generate text data in *Roth* in accordance with the shared language model and the first model for the first recognizer (limitation [1b]), and with the shared language model and the second model for the second recognizer (limitation [1c]), in have better recognition accuracy, a better ability to handle sentences, and a more time efficient means to provide a correct result for both recognizers.

147. In my opinion, a POSITA would have found an obvious platform for this implementation because *Roth* already describes using additional ranking methodologies to improve its recognition accuracy. *Roth*, 14:26-27 (“Any of the above-mentioned embodiments may be improved upon by employing ranking methodologies”), 15:4-5 (“When ranking methodologies are used, a higher confidence result can be obtained”); *see generally id.*, 14:26-17:2. Accordingly, *Roth*’s correlator can use the bigram model as a ranking methodology to modify *Roth*’s rankings of match words to obtain “a higher confidence result.” *See, e.g., id.*, 7:26-28, 14:26-27, 15:4-5. For example, *Roth*’s correlator can use the bigram model to increase or decrease the “confidence level weighting value” (*see, e.g., id.*, 15:4-24) assigned to each match word.

148. As to form, in my opinion, the addition of *Roberts*'s bigram model into *Roth*'s system could have been easily implemented as a separate input to *Roth*'s correlation step such that after a successful recognition of a first word, the recognized word is saved in memory for use in the correlation step for recognizing the subsequent second word in consultation with the bigram model and both recognition inputs. After each successfully corrected recognition, the correlation would inform the handwriting and speech recognition components back an indication of the correct word. Based on the correct word, the recognizers can re-train or adapt their internal representations of the words to be recognized.

149. Although *Roth* does not provide details of the feature extraction in box 8, it was well known that recognition in box 8 can be performed using Hidden Markov Models and acoustic models to model the extracted acoustic features that correlate to spoken phonemes. *See generally, e.g., Rabiner & Juang* section 6.13.3, page 376; *Rabiner; Klevans; Jelinek; Rabiner & Juang; Bailey*, 1:31-35, 5:62-6:17; 8:2-36; 11:9-37. *Roberts* then discloses a technique for those Hidden Markov Models to be updated, and that would be performed after the corrected word recognition, and performed within the feature extraction box, which is where the Hidden Markov Models would be stored.

150. Similarly, although *Roth* does not provide detail on the handwriting recognition, it was well known at the time that various pen strokes can be

recognized using Hidden Markov Models. *See, generally*, Ex. 1016, *Bellegarda et al.* “A Fast Statistical Mixture Algorithm for On-Line Handwriting Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 12, 1994, 1227-1233; *Bailey*, 1:31-35, 5:62-6:17; 8:2-36; 10:45-58; 11:9-37. A POSITA would have found it obvious to tune the models, in view of *Roberts*, based on the letters in the word that would have been recognized, in a manner similar to how the HMMs are tuned for speech, as disclosed in *Roberts. Id.*

151. In my opinion, a POSITA would also have found it obvious and beneficial to incorporate the training method of bigram models, as demonstrated in *Roberts*, to train the bigram model which would be incorporated into *Roth*. For example, *Roth* already discloses a method of user correction of misrecognitions. *See, e.g., Roth*, 8:15-19 (“in the (hopefully rare) case that the match was incorrect, the user will so indicate, either by touchpad input or by speech input to the system”). Because *Roth*’s method of user correction is compatible with *Roberts*’s method of user correction (*Roberts*, 4:16-20 (“The system also enables the user to select an individual word from a given indicated best matching word sequence, and to correct that word sequence by correcting the selected word. Preferably the word to be corrected can be selected by voice”)), the same method could be used to train the additional bigram model in *Roth*. Moreover, a POSITA would have expected similar benefits offered by *Roberts*’s trainable bigram model in *Roth*’s system. For

example, *Roberts* explains that its training method allows a user to “begin productive work on large vocabulary, natural language dictation tasks with little or no advance training of word models” (*Roberts*, 2:10-12), “build acoustic models from tokens that are collected during actual use of the system rather than during a separate training or enrollment process” (*id.*, 2:19-22), “easily correct errors made in continuous speech recognition” (*id.*, 2:24-25), and “develop a language model for use in a speech recognition system which is customized to his or her particular dictation, without requiring much additional effort to obtain such customization” (*id.*, 2:27-30). Accordingly, I believe that a POSITA would have been motivated to use *Roberts*’s method of training the additional bigram modal in *Roth*’s system during normal use to reap the same benefits including being able to perform “productive work on large vocabulary, natural language dictation tasks with little or no advance training of word models” (*id.*, 2:10-12), “build acoustic models from tokens that are collected during actual use of the system rather than during a separate training or enrollment process” (*id.*, 2:19-22), “easily correct errors made in continuous speech recognition” (*id.*, 2:24-25), and “develop a language model for use in a speech recognition system which is customized to his or her particular dictation, without requiring much additional effort to obtain such customization” (*id.*, 27-30).

152. Alternatively or additionally, I believe that a POSITA would have found it obvious to use *Roth*'s method of user correction to train the shared language models, e.g., feature templates used in its pronunciation-by-rule correlation method or the speech templates in the template store, to improve the accuracy of its recognizers, in view of *Roberts*'s method of updating a language model through user corrections. For example, when a user corrects a recognition output (*Roth*, 8:16-19), the system may use the correction data to change the pronunciation rules to account for differences in regional spelling or pronunciation.

153. For example, in British English, "color" is spelled COLOUR. In American English, it is spelled COLOR. If a user correction indicated that the person entered COLOUR instead of COLOR by handwriting, then the system could be trained to automatically switch to the British English roll of pronunciation such that the pronunciation-by-rule language model would be changed as a result of the correction. As another example, British English has five syllables for ALUMINUM, but American English has only four. A user's correction of a misrecognition might update the word-to-letter and word-to-sound rules within the language model. In another example, the word CHOCOLATE can be pronounced with two syllables or three syllables by different users, but a given speaker will consistently pronounce it one way. So, the speaker's correction can update the two-syllable version of the word to the three-syllable version of the word in the

language model such that the model becomes adapted to that speaker. In these ways, the combination would train the *shared* language model so that it provides more accurate correlations between spoken input and pen input, which in turn generates more accurate word spellings.

2. Claim 2

“The system of claim 1 wherein:

(a) the first type of message recognition is speech recognition, the first type of message input is voice input, and the first model is an acoustic model; and

(b) the second type of message recognition is handwriting recognition, the second type of message input is freehand input, and the second model is a handwriting model.”

154. In my opinion, *Roth* discloses these additional limitations. For example, *Roth* discloses that the first type of message recognition is speech recognition, with the first type of message input being voice input, and the first model is an acoustic model. *See, e.g., Roth*, 1:1-4 (“This invention relates generally to apparatus for receiving handwritten (calligraphic) input, and relates more particularly to a novel technique for augmenting the handwritten input with audio (spoken) input”), 13:8-14 (“Those skilled in the art will appreciate that while the embodiments have been disclosed with respect to one particular way of characterizing a spoken input (extraction of acoustic features and comparison with a set of acoustic-feature templates), the inventive aspects of the invention do not

depend on the use of such templates”), 11:17-21 (“[t]he invention may be optionally embodied with the opposite sequence, namely (1) analyzing the spoken input to derive a number of candidate speech portions (e.g. words) and (2) receiving written input to disambiguate (e.g. narrow down the candidates to one result)”). *See also* discussion above for elements [1a] and [1b] in Section IX.C.1.b and IX.C.1.c.

155. Furthermore, *Roth* teaches that the second type of message recognition is handwriting recognition, with the second type of message input being handwriting input. *See, e.g., Roth*, 5:14-6:2 (“Known recognition methodologies which have been employed include feature matching, in which features of the to-be-recognized character are tested for match with characters in a feature database; time sequence of zones, directions, or extremes, where a simple database match, or a binary tree analysis, may be employed regarding relatively discrete information about the character-formation temporal data; curve matching, in which the curves making up the character are compared with standard curves in a database; or stroke codes, where parts of a character are classified and enumerated, then compared with a database ... the recognition is performed, but the recognition algorithm only requires that the range of possible recognition results be confined to a small number of matches”), 11:3-8 (“It will be noted that the above-described embodiment is disclosed with respect to one particular sequence, namely

(1) analyzing the handwritten input to derive a number of candidate writing portions (e.g. words) and (2) receiving spoken input to disambiguate (e.g. narrow down the candidates to one result)”). *see also* discussion for elements [1a] and [1c] in Ground 5). It would also have been obvious, as I noted above, for the second model to be a handwriting model.

3. Claim 3

“The system of claim 2 further comprising a stylus, wherein the shared language model is trainable responsive to user correction of misrecognition by the first message recognizer by (a) movement of the stylus across visual indicia of dictated text that has been erroneously generated to specify a selection from the text, and (b) correction of the selection.”

156. In my opinion, *Roth*, alone or with *Roberts*, discloses or at least renders obvious this additional limitation. *Roth* discloses using a stylus to correct misrecognition by a recognizer by (a) movement of the stylus across visual indicia of dictated text that has been erroneously generated to specify a selection from the text, and (b) correction of the selection, which a POSITA would have understood to include misrecognition by the speech recognizer. For example, *Roth* teaches that “[t]ypically an input device such as a touchpad is employed, which receives user input by a stylus (or finger)” (*Roth*, 1:7-10), and that “[i]n the (hopefully rare) case that the match was incorrect, the user will so indicate, either by touchpad input or by speech input to the system” (*id.*, 8:8:15-19). *Roth* also teaches that “the recognition result is provided to the user on a visual display, and an opportunity is

provided so the user may confirm the correctness of the recognition.” *Id.*, 3:12-15.

Accordingly, I believe that a POSITA would have understood that, because *Roth* receives user input through a stylus, *Roth* also receives user correction input through a stylus. A POSITA would also have understood that the stylus, in correcting the recognition result, moves across visual indicia of dictated text that has been erroneously generated on a tablet (e.g., “recognition result”).

157. Moreover, *Roberts* teaches a system with a shared language model trainable responsive to user correction of misrecognition by the first message recognizer (e.g., speech recognizer) by either a user’s speech or keyboard input correction. *See supra* Section IX.A.1.e and IX.C.1.e. Thus, a POSITA would have found it obvious to train the shared model in the system of *Roth* and/or modified in view of *Roberts* (element [1d] above) responsive to user correction of misrecognition in the manner described in *Roth*, e.g., using a stylus. *See, e.g., Roberts*, 4:15-20 (“The system also enables the user to select an individual word from a given indicated best matching word sequence, and to correct that word sequence by correcting the selected word. Preferably the word to be corrected can be selected by voice.”), Abstract (“When the correct word is chosen by the user, the speech sample is used to create or update an acoustic model for the word, without further intervention by the user. As the system is used, it also constantly

updates its statistical language model. The system gets more and more word models and keeps improving its performance the more it is used”).

158. In my opinion, this would have been especially obvious because *Roberts* also teaches that the correction in *Roberts* can be performed using different modalities, such as speech or keyboard. *Id.*, 2:48-55 (“Preferably the menu of character sequences is a list of best matching words produced by speech recognition of one of the training utterances, and the user can select and edit the desired menu choice by voice. It is also preferred that the system is used as part of a speech recognition system which produces as its output for each spoken word either the word recognized or the word edited and selected for that spoken word”), Abstract (“By keyboard or voice command, the user may choose the correct word from the list or may choose to edit a similar word if the correct word is not on the list”).

4. Independent Claim 7

- a. [7P] “*A method for performing message recognition with a shared language model, the method comprising:*”**

159. Independent claim 7 is substantially identical to claim 1, except it is directed to a method. As discussed in limitations [1P] and [1a], I believe that *Roth* in view of *Roberts* renders obvious this limitation. Supra Sections IX.C.1.a and IX.C.1.b.

- b. [7a] ***“(a) performing message recognition of a first type, responsive to a first type of message input, to provide text data in accordance with both the shared language model and a first model specific to the first type of message recognition;”***

160. This limitation is substantially identical to limitations [1a] and [1b].

For the reasons discussed for [1a] and [1b], I believe that *Roth* in view of *Roberts* renders obvious this limitation. Supra Sections IX.C.1.b and IX.C.1.c.

- c. [7b] ***“(b) performing message recognition of a second type, responsive to a second type of message input, to provide text data in accordance with both the shared language model and a second model specific to the second type of message recognition; and”***

161. This limitation is substantially identical to limitations [1a] and [1c].

For the reasons discussed for [1a] and [1c], I believe that *Roth* in view of *Roberts* renders obvious this limitation. Supra Sections IX.C.1.b and IX.C.1.d.

- d. [7c] ***“(c) training the shared language model responsive to user correction of error in message recognition of either of the first and second types, thereby improving accuracy of each of the first and [sic] second types of message recognition.”***

162. This limitation is substantially identical to limitations [1a] and [1d].

For the reasons discussed for limitations [1a] and [1d], I believe that *Roth* in view of *Roberts* renders obvious this limitation. Supra Sections IX.C.1.b and IX.C.1.e.

5. Claim 8

“The method of claim 7 wherein:

(a) performing the first type of message recognition comprises performing speech recognition responsive to voice input in accordance with the shared language model and a first model that is an acoustic model; and

(b) performing the second type of message recognition comprises performing handwriting recognition responsive to freehand input in accordance with the shared language model and a second model that is a handwriting model.”

163. This claim is substantially identical to claim 2. For the reasons discussed above for claim 2, I believe that *Roth* in view of *Roberts* renders obvious this limitation. Supra Section IX.C.2.

6. Claim 9

“The method of claim 8 wherein training the shared language model includes (a) moving a stylus across visual indicia of dictated text that has been erroneously generated to specify a selection from the text, and (b) correcting the selection.”

164. This claim is substantially identical to claim 3. For the reasons discussed above for claim 3, I believe that *Roth* in view of *Roberts* renders obvious this limitation. Supra Section IX.C.3.

**D. Claims 4, 5, 10, and 11 Would Have Been Obvious over *Roth*,
Roberts, and *Verrier* (Ground 6)**

1. Claim 4

“The system of claim 3 wherein:

(a) the stylus includes a pointing interface and a microphone;

(b) the system further comprises a tablet responsive to the pointing interface;

(c) the voice input is from the microphone; and

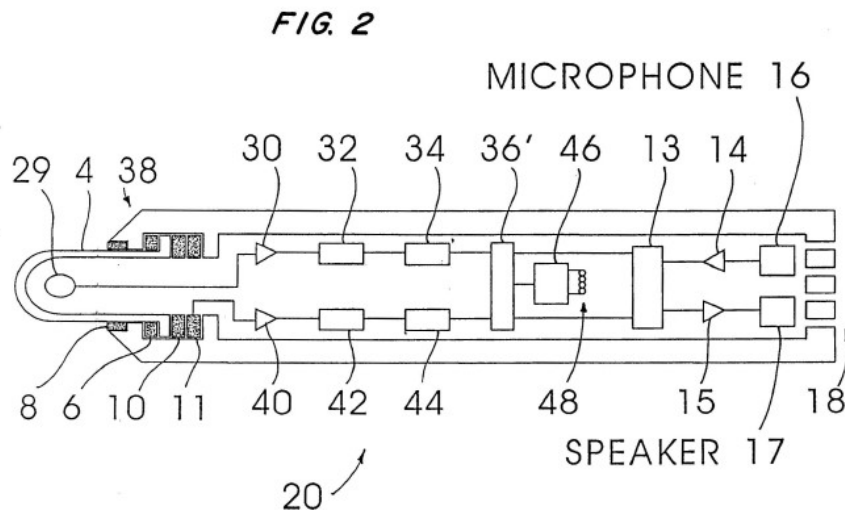
(d) the freehand input is from the pointing interface being manually manipulated on the tablet.”

165. In my opinion, *Roth* in view of *Roberts* and *Verrier* renders these limitations obvious.

166. First, *Roth* discloses “*stylus includes a pointing interface*” and “*freehand input is from the pointing interface being manually manipulated on the tablet,*” for it discloses using a stylus for freehand input on a touchpad. *Roth*, 1:7-10 (“an input device such as a touchpad is employed, which receives user input by a stylus (or finger) and provides information about the position of the stylus in an orthogonal reference frame”); *see also Roth*, claim 3. A POSITA would have understood this is an input from the pointing interface of a stylus being manually manipulated on the tablet.

167. Although *Roth* does not expressly disclose that the “*stylus includes ... a microphone*” or “*the voice input is from the microphone,*” incorporating a

microphone in the pen of *Roth* and providing voice input from the microphone was known to a POSITA at the time. As explained above, *Verrier* discloses a stylus with a microphone for receiving voice input. See, e.g., *Verrier*, Abstract (“microphone and speaker are included in an untethered stylus ... to enable multimedia communication between the user and the pen-based computer system, through a digitizing display”), 18:35-40, Fig. 2. *Verrier*’s stylus is shown below in Figure 2:



168. *Id.*, Fig. 2. It would have been obvious to a POSITA to incorporate a microphone in *Roth*’s pen and receive voice input from the microphone for the acquisition of high-quality speech in a pen-based system to accommodate voice communications at the user stylus interface. *Verrier*, 3:27-35. Indeed, a POSITA would have known that placing the microphone in the stylus improves the quality of the audio signal due to the user’s proximity to the stylus while dictating

regardless of the location of the computer device, which in turn improves speech recognition. Even the '405 patent admits “[i]t has long been known to incorporate a microphone into a stylus for the acquisition of high-quality speech in a pen-based” system. '405 patent, 2:54-64. A POSITA would have known that incorporating a microphone into *Roth*'s stylus would have been the simple substitution of known elements according to their known functions to achieve predictable results. A POSITA would have been motivated to minimize overall hardware and power consumption while increasing system portability. The POSITA would have had a reasonable expectation of success in doing so.

169. Finally, *Roth* discloses “***a tablet responsive to the pointing interface.***” *Roth*, Fig. 1, element 1; 2:35-3:2 (“user enters a speech portion (a word, a part of a word, or multi-word phrase) by writing it on a tablet and speaking it into a microphone”), 8:30-33 (“touchpad tablet 1 with interface 22 plugged into an expansion slot of bus 27, microphone 6 with speech recognition (feature extraction) interface 24, and an optional speaker with speech synthesis card 26”).

2. Claim 5

“The system of claim 4 wherein the visual indicia of dictated text appears on the tablet.”

170. In my opinion, *Roth* in view of *Roberts* and *Verrier* renders this obvious. As I explained in connection with claims 3 and 4 above, *Roth* discloses the use of a tablet as a human-computer interface and its display of the visual

indicia of dictated text. *See, e.g., Roth*, 3:12-15 (“Optionally a speech synthesizer speaks the word, or the recognition result is provided to the user on a visual display, and an opportunity is provided so the user may confirm the correctness of the recognition”), 8:11-19 (“Instead of or in addition to an audio synthesis of the match character, there may be a visual display of the match character as indicated by block 28. Assuming the user agrees with the match, the process continues with the next handwriting to be recognized. As with the audio synthesis technique just described, in the (hopefully rare) case that the match was incorrect, the user will so indicate, either by touchpad input or by speech input to the system”).

171. Between *Roth*’s priority date in 1990 and the ’405 patent’s filing date in 2002, the fast and commonplace development of touch-sensitive screens in portable devices (e.g., tablets) made it obvious that the integration of various functions, including touch input and display, into a single integrated device, rather than two distinct devices, was more desirable for both user-portability, power efficiency, and manufacturing cost. Thus, I believe that a POSITA would have found it obvious to integrate *Roth*’s input tablet and output display into a single device, such that the visual indicia of dictated text appears on the tablet, to at least improve portability, reduce power expenditures, and reduce manufacturing costs.

3. Claim 10

“The method of claim 9 wherein:

(a) performing speech recognition responsive to voice input includes responding to voice input from a microphone in the stylus; and

(b) performing handwriting recognition responsive to freehand input includes responding to manual manipulation of the stylus on a tablet.”

172. For the reasons discussed above for claims 1, 4, and 7, I believe that *Roth* in view of *Roberts* and *Verrier* renders obvious the additional limitations of claim 10. Supra Section IX.C.1, IX.D.1, and IX.C.4.

4. Claim 11

“The method of claim 10 wherein the visual indicia of dictated text appears on the tablet.”

173. This claim is substantially identical to claim 5. For the reasons I discussed above for claim 5, I believe that *Roth* in view of *Roberts* and *Verrier* renders obvious this additional limitation. Supra Section IX.D.2.

E. Claims 6 and 12 Would Have Been Obvious over *Roth*, *Roberts*, and *Vo* (Ground 7)

1. Claim 6

“The system of claim 1 wherein the shared language model comprises a syntactic model and a semantic model.”

174. In my opinion, *Roth* in view of *Roberts* and *Vo* renders obvious the shared language model comprises “*a syntactic model*” because a POSITA would have understood the disclosed bigram models describe the probabilities of words

appearing in sequence, namely, syntactic information. *See, e.g., Roberts*, 13:50-53 (describing word bigrams); *see also supra*, Ground 5, [1d]. Although *Roth* in view of *Roberts* does not teach that its shared language model comprises a semantic model, the use of a semantic model was well known in the art to improve recognition. *Vo*, for example, discloses the use of a semantic model as a language model. *See, e.g., Vo*, 3546 (“The text string output from the speech recognizer is processed by the Phoenix semantic parser ... The parser compiles a grammar specifying semantically meaningful fragments of text into an efficient recurrent transition network that identifies these fragments in the recognized utterance. It is capable of skipping unknown words and unmatched fragments and can therefore deal gracefully with ungrammatical sentences. The concepts identified by the parser serve to determine which slots of the speech frame should be filled and what scores should be assigned to the slot values.”). Even the ’405 patent admits that the use of semantic and syntactic models in recognizers were well known in the art. *See, e.g., ’405 patent*, 26:11-21 (“semantic model 354 may be of the type disclosed in U.S. Pat. No. 5,828,999 to Bellegarda et al.” and “a language model may be of the type disclosed in U.S. Pat. No. 5,839,106, issued Nov. 17, 1998 to Bellegarda and U.S. Pat. No. 5,828,999, issued Oct. 27, 1998 to Bellegarda et al”); *see also Ex1029*, 172-175 (describing syntactic and semantic models). Thus, a POSITA would have found it obvious to incorporate a semantic model in *Roth* to improve

upon its recognition accuracy and “deal gracefully with ungrammatical sentences.”

Vo, 3546; *see also* discussion of [1d] in Grounds 1 and 2, Section IX.A.1.e. And a

POSITA would have had a reasonable expectation of success in doing so.

2. Claim 12

“The method of claim 7 wherein performing message recognition in accordance with the shared language model comprises performing message recognition in accordance with a syntactic model and a semantic model.”

175. This claim is substantially identical to claim 6. For the reasons discussed for claim 6, *Roth* in view of *Roberts* and *Vo* renders obvious this additional limitation. *See supra* Section IX.E.1.

X. Conclusion

176. For the reasons set forth above, I believe claims 1-12 of the ’405 patent would have been obvious in view of the prior art.

177. In signing this Declaration, I understand it will be filed as evidence in a contested case before the Patent Trial and Appeal Board of the United States Patent and Trademark Office. I understand I may be subject to cross-examination in this case, and that any cross-examination will take place in the United States. If cross-examination is required of me, I will appear for cross-examination within the United States during the time allotted for cross-examination.

178. I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and

Declaration of Christopher Schmandt
U.S. Patent No. 6,904,405

that those statements were made with the knowledge that willful false statements
and the like so made are punishable by fine, imprisonment, or both, under Section
1001 of Title 18 of the United States Code.

Respectfully submitted,

Date: September 12, 2023

By: 

Christopher Schmandt