UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

SAMSUNG ELECTRONICS CO., LTD., and
SAMSUNG ELECTRONICS AMERICA, INC.,

Petitioners

v.

CERENCE OPERATING COMPANY,

Patent Owner

Case IPR2025-00460

U.S. Patent No. 8,914,291

DECLARATION OF CHRISTOPHER M. SCHMANDT IN SUPPORT OF PETITION FOR *INTER PARTES* REVIEW OF U.S. PATENT NO. 8,914,291

Samsung v. Cerence
IPR2025-00460
Exhibit 1002

TABLE OF CONTENTS

I. INTRODUCTION 5

II. BACKGROUND AND QUALIFICATIONS 6

III. MATERIALS REVIEWED 11

IV. LEVEL OF ORDINARY SKILL IN THE ART 13

V. RELEVANT LEGAL STANDARDS 15

VI. SUMMARY OF OPINIONS 19

VII. OVERVIEW OF THE '291 PATENT 19

VIII. OVERVIEW OF THE PRIMARY PRIOR ART REFERENCES 24

A. Analogous Art 24

B. *Walker* 26

C. *Matsumoto* 30

IX. CLAIM CONSTRUCTION 34

X. SPECIFIC GROUNDS FOR CHALLENGE 35

XI. DETAILED EXPLANATION OF GROUNDS 36

A. Ground 1A: *Walker* in Combination With *Matsumoto* And *Malsheen* Renders Obvious Claims 1-4, 6-11, 13-17, and 19-20 36

1. Independent Claim 1 36

a. 1[pre]: A method for providing speech output for a speech-enabled application, the method comprising: 36

b. 1[a]: receiving from the speech-enabled application a text input comprising a text transcription of a desired speech output; 39

c. 1[b]: identifying at least one first portion of at least one token of the text input that differs from at least one corresponding first portion of at least one other token of the text input, and at least one second portion of the at least one token that does not differ from at least one corresponding second portion of the at least one other token; 43

(i) "Token" as used in the '291 claims and specification 44

(ii) Motivation to Combine *Walker* and *Matsumoto* ...47

(iii) Example text weather report that a person of ordinary skill in the art would have understood to be taught by the *Walker-Matsumoto* combination 54

(iv) Motivation to combine *Walker-Matsumoto* with *Malsheen* 59

(v) Application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report 63

d. 1[c]: assigning contrastive stress to the identified at least one first portion of the at least one token, but not to the identified at least one second portion of the at least one token; 71

e. 1[d]: generating, using at least one computer system, an audio speech output corresponding to at least a portion of the text input, the audio speech output comprising at least one portion carrying contrastive stress corresponding to the at least one first portion of the at least one token of the text input, to contrast with at least one other portion of the audio speech output corresponding to the at least one corresponding first portion of the at least one other token of the text input; and 86

f. 1[e]: providing the audio speech output for the speech-enabled application. 90

2. Independent Claim 8 91

a. 8[pre]: Apparatus for providing speech output for a speech-enabled application, the apparatus comprising: .91

b. 8[a]: a memory storing a plurality of processor-executable instructions; and 8[b]: at least one processor, operatively coupled to the memory, that executes the instructions to: 91

c. 8[c]-[g] 94

3. Independent Claim 14 94

a. 14[Pre]: At least one non-transitory computer-readable storage medium encoded with a plurality of computer-executable instructions that, when executed, perform a method for providing speech output for a speech-enabled application, the method comprising: 94

b. 14[a]-[e] 99

4. Dependent Claims 2-4, 6-7, 9-11, 13, 15-17 and 19-20 99

a. Claim 2.....99

(i) 2[a]: The method of claim 1, wherein the identifying comprises: identifying a plurality of tokens of the text input of a same text normalization type for which a contrastive stress pattern is to be applied; 99

(ii) 2[b]: identifying the at least one token to be rendered with contrastive stress from among the plurality of tokens of the same text normalization type.....103

b. Claim 3: The method of claim 2, wherein the same text normalization type is a text normalization type selected from the group consisting of: an alphanumeric sequence type, an address type, a Boolean value type, a currency type, a date type, a digit sequence type, a fractional number type, a proper name type, a number type, an

ordinal number type, a telephone number type, a flight number type, a state name type, a street name type, a street number type, a time type and a zipcode type.....106

c. Claim 4: The method of claim 2, wherein the plurality of tokens are identified based at least in part on at least one indication in the text input that the contrastive stress pattern is desired in association with the plurality of tokens. 106

d. Claim 6.....108

(i) 6[a]: The method of claim 2, wherein identifying the plurality of tokens comprises: tokenizing the text input; 108

(ii) 6[b]: automatically identifying the text normalization type of the plurality of tokens; and 6[c] automatically determining that the contrastive stress pattern is to be applied for the plurality of tokens.....113

e. Claim 7: The method of claim 2, wherein the at least one token to be rendered with contrastive stress is identified based at least in part on an order of the plurality of tokens in the text input. 114

f. Claims 9-11, 13.....116

g. Claims 15-17 and 19-20 116

A. Ground 1B: *Walker* in Combination With *Matsumoto* and *Bellegarda* Renders Obvious Claims 1-4, 6-11, 13-17, and 19-20.....117

1. 1[c], 8[e], 14[c] 117

2. All Other Features of Claims 1-4, 6-11, 13-17, and 19-20.....121

XII. SECONDARY CONSIDERATIONS.....123

I. INTRODUCTION

1. I have been retained by Samsung Electronics Co., Ltd. and Samsung Electronics America, Inc. (collectively, "Samsung" or "Petitioners") as an independent expert consultant in this proceeding before the United States Patent and Trademark Office ("PTO").

2. I am being compensated at a rate of $550/hour for my services in this proceeding, which is my regular and customary rate.

3. My compensation is in no way contingent on the nature of my finding, the presentation of my findings in testimony or this declaration, or the outcome of this or any other proceeding. I have no other interest in this proceeding.

4. I have been asked to consider whether certain references disclose and/or suggest the features recited in claims 1-4, 6-11, 13-17, and 19-20 of U.S. Patent No. 8,914,291 ("the '291 patent") (Ex. 1001).[1] My opinions are set forth below.

[1] Where appropriate, I refer to exhibits which I understand will be attached to the petition for *inter partes* review of the '291 patent (the "Petition").

II. BACKGROUND AND QUALIFICATIONS

5. I am an independent consultant. All of my opinions stated in this declaration are based on my own personal knowledge and professional judgment. In forming my opinions, I have relied on my knowledge and extensive experience in the field of compute-based speech recognition.

6. I am over 18 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. A copy of my current curriculum vitae, which details my education and professional experience, is included as Ex. 1003 in this proceeding. The following provides an overview of some of my experience that is relevant to the matters set forth in this declaration.

7. I have extensive experience with computer-based speech synthesis technology, which provides me with an understanding of the subject matter described and claimed in the ’291 patent.

8. I retired five years ago after a 40 year career at the Massachusetts Institute of Technology (“MIT”); for which 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.

9. 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 since 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 multi-media and user interfaces, with a strong focus on speech-based systems and portable devices, some combined with location-based services. I have over 70 published conference and journal papers and one book in the field of speech technology and user interaction.

10. 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.

11. 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").

12. 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 Ph.D. level. I oversaw the Masters and Ph.D. 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.

13. I was a pioneer in developing voice-based user interfaces, starting with "Put That There" in 1980, which was one of the first systems employing multi-modal input (speech and pointing) as well as conversational ability (hearing speech input and responding with synthetic speech output). Put That There used both vocal tract model synthesis and concatenative synthesis to maintain a conversation to understand spoken commands, which sometimes included mis-recognized words.

14. Further, I built a number of systems, starting in 1983, to read electronic mail and other documents over the telephone; to do so intelligibly required analyzing the text to find patterns such as dates, times, and email addresses, and applying specialized pronunciation rules to these phrases once detected. To extend such systems to reading both general and financial news, I

extended these rules to properly pronounce acronyms (e.g., “IBM” is spoken as three letters, not a single word) and built a lexicon of company names and technical terms, to provide specialized substitute pronunciations. Similar concepts (i.e. detecting word patterns indicative of particular patterns of text) were also applied to reading stock quotes, telephone access to electronic calendars, and speaking data from a personal contact list.

15. One of the most demanding uses of synthetic speech was Back Seat Driver, a pioneering (1988) system which provided real time spoken driving directions in a moving automobile. Much like the systems which are commonplace today, Back Seat Driver had to determine a route from a map database, and then speak the route, at just the right times, in a way which reflected the real street geometry, including forced turns, T intersections, merges, multi-lane roads, and traffic rotaries. To help maintain clarity, it was important to differentiate when information spoken previously was being repeated (“At the next intersection turn left” and later “Turn left at this intersection.”). The best solution to this problem came in the form of collaboration with AT&T Bell Labs by means of a student project applying Back Seat Driver text and rules to their speech synthesizer which was better equipped to generate contrastive stress.

16. Other work, sponsored by DARPA (1986) and NSF (1995), explored detection of discourse structure (in particular, the introduction of new information) in human speech, and applying that to synthetic speech. Our early work used concatenative speech (Linear Predictive Coding) due the relatively easy manipulation of its pitch contours; this offered an easy although less flexible testbed for experimenting with contrastive stress and other forms of pitch accents and emphasis. Follow-on work used general synthesis through use of a vocal tract model to create synthetic speech reflecting affect or emotional state (boredom, fright, anger, joy), which required stressing different words depending on affective state, and in doing so to manipulate pitch, duration, and amplitude of the synthetic phonemes.

17. An additional example of related work included an MS thesis project (Jordan Slott, 1996) which described and implemented a text-to-speech markup language, and a number of "experts" which were modules to analyze text for particular patterns, such as telephone numbers, names, computer jargon terms, and compound nouns, and insert markup text tags into the text stream to be synthesized, noting where these patterns have been identified. Later, during synthesis of the text, when a tag is detected the corresponding processing expert is invoked to speak the type of text found within the tags. This work provided a more

generalized and extensible framework, including tags for prosodics, than the earlier work described above for correct pronunciations reading email and news.

III. MATERIALS REVIEWED

18. The opinions contained in this declaration are based on the documents I reviewed, my professional judgment, as well as my education, experience, and knowledge regarding computer-based speech synthesis technology.

19. In forming my opinions expressed in this declaration, I reviewed the following materials:

- U.S. Patent 8,914,291 ("*'291Pat*") (Ex. 1001);
- File History of U.S. Patent 8,914,291 ("*'291FH*") ("Ex. 1004");
- U.S. Patent App. No. 2001/0049602 ("*Walker*") (Ex. 1005);
- U.S. Patent App. No. 2007/0233492 ("*Matsumoto*") (Ex. 1006);
- U.S. Patent 5,634,084 ("*Malsheen*") (Ex. 1007);
- W3C Speech Synthesis Markup Language (SSML) Ver. 1.1 WD4 , September 4, 2007 ("*W3C*") (Ex. 1008);
- U.S. Patent 7,010,489 ("*Lewis*") (Ex. 1009);
- U.S. Patent 7,177,807 ("*Schmid*") (Ex. 1010);
- U.S. Patent 6,865,533 ("*Addison*") (Ex. 1011);
- U.S. Patent 7,725,321 ("*Bennett*") (Ex. 1012);

- File History of U.S. Patent 8,447,610 ("*'610FH*") (Ex. 1013);
- U.S. Patent 7,313,523 (*"Bellegarda"*) (Ex. 1014);
- U.S. Provisional Application No. 60/205,000 ("*'000Prov.*") (Ex. 1015);
- *Speech Synthesis*, in *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Daniel Jurafsky & James H. Martin Oct. 13, 2007) ("*Jurafsky*") (Ex. 1016);
- U.S. Patent 8,682,671 ("*'671Pat.*") (Ex. 1017);
- U.S. Patent 8,447,610 ("*'610Pat*") (Ex. 1018);
- Redline comparison of the specification of the '291 patent to the specification of the '610 patent (Ex. 1019)
- Summary of patents/applications in the '291 patent's family (Ex. 1022);

and any other materials I refer to in this declaration in support of my opinions.

20. 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 appreciation of how a person of ordinary skill in the art would have understood the claims and the specification of the '291 patent at the

time of the alleged invention, which I have been asked to consider as early as February 12, 2010. My opinions reflect how one of ordinary skill in the art would have understood the '291 patent, the prior art to the patent, and the state of the art at the time of the alleged invention.

21. It is my opinion that certain references disclose and/or suggest all the features recited in claims 1-4, 6-11, 13-17, and 19-20 ("Challenged Claims") of the '291 patent, as I discuss in detail below.

IV. LEVEL OF ORDINARY SKILL IN THE ART

22. I have been informed and understand that, in the context of an invalidity analysis, a person having ordinary skill in the art is a hypothetical person who looks to prior art at the time of the invention. I further understand that the factors that may be considered in determining the level of ordinary skill include: (1) the problems encountered in the art; (2) the prior art solutions to the problems encountered in the art; (3) the rapidity of innovation; (4) the sophistication of the technology; and (5) the education level of individuals actively working in the field. I understand that these factors need not all be considered for the analysis and that one or more of these factors may control.

23. I was asked to provide my opinion on the level of one of ordinary skill in the art with respect to the alleged invention of the '291 patent by February 12,

2010.[2] Based on my consideration of the factors above, I believe a person of ordinary skill in the art would have had a Bachelor of Science (or equivalent) in electrical engineering or computer science, and two years of work experience in the field of speech and natural language processing, including text-to-speech systems. A greater amount of education could compensate for less work experience, and vice versa.

24. As of February 12, 2010, I met, and in fact exceeded, the qualifications of a person of ordinary skill in the art. To be clear, all of my opinions in this declaration are from the perspective of one of ordinary skill in the art as I have defined it here during the relevant timeframe.

[2] I have been informed that Petitioners contend that the earliest priority date to which the claims of the '291 patent are possibly entitled is August 9, 2010. I have not been asked to opine on this issue. I note, however, that my opinion regarding the qualifications of a person of ordinary skill in the art would be the same regardless of which of the two priority dates the '291 patent is entitled.

V. RELEVANT LEGAL STANDARDS

25. I am not an attorney and offer no legal opinions, but in the course of my work, I have had experience studying and analyzing patents and patent claims from the perspective of a person skilled in the art.

26. For the purposes of this declaration, I have been informed about certain aspects of the law that are relevant to forming my opinions. My understanding of the law is as follows.

27. Petitioners' counsel has informed me that a published nonprovisional application is entitled to claim priority to the filing date of its provisional application (as opposed to the filing date of the published non-provisional application) if (1) the provisional application discloses the portions of the published non-provisional application relied upon as prior art, and (2) at least one claim of the published non-provisional claim is supported by the provisional application, i.e., the provisional application must disclose the subject matter of the claim and enable a person of ordinary skill in the art to make and use the claim.

28. Petitioners' counsel has informed me that for the prior art to inherently disclose a claimed limitation, the prior art need not expressly disclose the limitation, so long as the claimed limitation necessarily flows from a disclosure in the prior art. I also understand that it is acceptable to examine evidence outside

the prior art reference (extrinsic evidence) in determining whether a feature, while not expressly discussed in the reference, is necessarily present in that reference.

29. Petitioners' counsel has informed me that a patent claim can be considered to have been obvious to a person having ordinary skill in the art at the time the application was filed. I am informed that this means that, even if all of the requirements of a claim are not found in a single prior art reference, the claim is not patentable if the differences between the subject matter in the prior art and the subject matter in the claim would have been obvious to a person having ordinary skill in the art at the time of the invention.

30. I have been informed by Petitioners' counsel that a determination of whether a claim would have been obvious should be based upon several factors, including, among others:

- the level of ordinary skill in the art at the time the application was filed;
- the scope and content of the prior art;
- what differences, if any, existed between the claimed invention and the prior art; and
- any "secondary indicia" of non-obviousness, if they are of record.

31. I have been informed by Petitioners' counsel that a single reference can render a patent claim obvious if any differences between that reference and the claims would have been obvious to a person having ordinary skill in the art. Alternatively, I understand that the teachings of two or more references may be combined in the same way as disclosed in the claims, if such a combination would have been obvious to one having ordinary skill in the art. In determining whether a combination based on either a single reference or multiple references would have been obvious, the following are examples of approaches and rationales that may be considered:

- whether the teachings of the prior art references disclose known concepts combined in familiar ways, and when combined, would yield predictable results;
- whether a person of ordinary skill in the art could implement a predictable variation, and would see the benefit of doing so;
- whether the claimed elements represent one of a limited number of known design choices, and would have a reasonable expectation of success by those skilled in the art;
- whether a person of ordinary skill would have recognized a reason to combine known elements in the manner described in the claim;

- whether there is some teaching or suggestion in the prior art to make the modification or combination of elements claimed in the patent; and
- whether the innovation applies a known technique that had been used to improve a similar device or method in a similar way.

32. I understand that "secondary indicia" of non-obviousness may include certain objective factors, such as: commercial success of products practicing the claimed invention; long-felt but unsolved need; teaching away; unexpected results; copying; and praise by others in the field. These factors are generally referred to as "secondary considerations" or "objective indicia" of non-obviousness. I understand, however, that for such objective evidence to be relevant to the (non)obviousness of a claim, there must be a causal relationship (called a "nexus") between the claim and the evidence and that this nexus must be based on a novel element of the claim rather than something in the prior art. I also understand that even when they are present, secondary considerations may be unable to overcome primary evidence of obviousness (such as motivation to combine with predictable results) that is sufficiently strong.

33. I have been informed by Petitioners' counsel that one of ordinary skill in the art has ordinary creativity and is not an automaton. Petitioners' counsel has

also informed me that in considering obviousness, obviousness may not be determined using the benefit of hindsight, including hindsight derived from the patent being considered.

VI. SUMMARY OF OPINIONS

34. For the reasons I discuss below, it is my opinion that claims 1-4, 6-11, 13-17, and 19-20 of the ’291 patent are rendered obvious by the prior art.

VII. OVERVIEW OF THE ’291 PATENT

35. The ’291 patent titled “Method and apparatus for generating synthetic speech with contrastive stress” relates generally to “[t]echniques for generating synthetic speech with contrastive stress.” Ex. 1001 (*’291Pat*), Abstract.

36. The ’291 patent aims to improve text-to-speech technology by adding stress where a human speaker would naturally apply stress in order to help the listener understand the speaker:

> Human speakers naturally apply contrastive stress to emphasize a word or syllable contrary to its normal accentuation, in order to contrast it with an alternative word or syllable or to focus attention on it.

Id., 8:20-23.

> ***Contrastive stress can be an important tool in human understanding of meaning as conveyed by spoken language***; however, conventional automatic speech synthesis technologies have not taken advantage of contrastive stress as an opportunity to improve intelligibility, naturalness and

> effectiveness of machine generated speech. Applicants have recognized that a primary focus of many automated information systems is to provide numerical values and other specific datums to users, who in turn often have preconceived expectations about the kind of information they are likely to hear. Information can often be lost in the stream of output audio when a large number of words must be output to collect necessary parameters from the user and to set the context of the systems response. ***Applicants have appreciated, therefore, that a system that can highlight that although the user expected to hear "this", the actual value is "that", may allow the user to hear and process the information more easily and successfully***.

Id., 8:32-48.[3]

37. For example, the '291 patent describes that in an utterance such as "not nineteen, but twenty," the "speaker may have particularly emphasized the number 'twenty' to distinguish and contrast it from the other number 'nineteen' in the utterance." *Id*., 12:38-44.

38. The '291 patent achieves this effect by first receiving "a text transcription of a desired speech output." *Id*., 4:45-46. A synthesis system "analyze[s] the text input to determine which portion(s) of the speech output should carry contrastive stress." *Id*., 9:5-8. Then, the synthesis system "may

[3] I have added all **bold**/*italics*/color emphases and annotations unless I note otherwise.

automatically identify which tokens (e.g., words) should contrast with each other." *Id*., 9:13-15. "[A]fter identifying which word(s) and/or syllable(s) should carry contrastive stress, the synthesis system may apply the contrastive stress to the identified word(s) and/or syllable(s) through increased pitch, amplitude[,] and/or duration[.]" *Id*., 9:23-27.

39. After this analysis and application of stress by the synthesis system, a computer system then generates "an audio speech output corresponding to at least a portion of the text input." *Id*., 4:46-48. The audio speech output contains "at least one portion carrying contrastive stress to contrast with at least one other portion of the audio speech output." *Id*., 4:48-50.

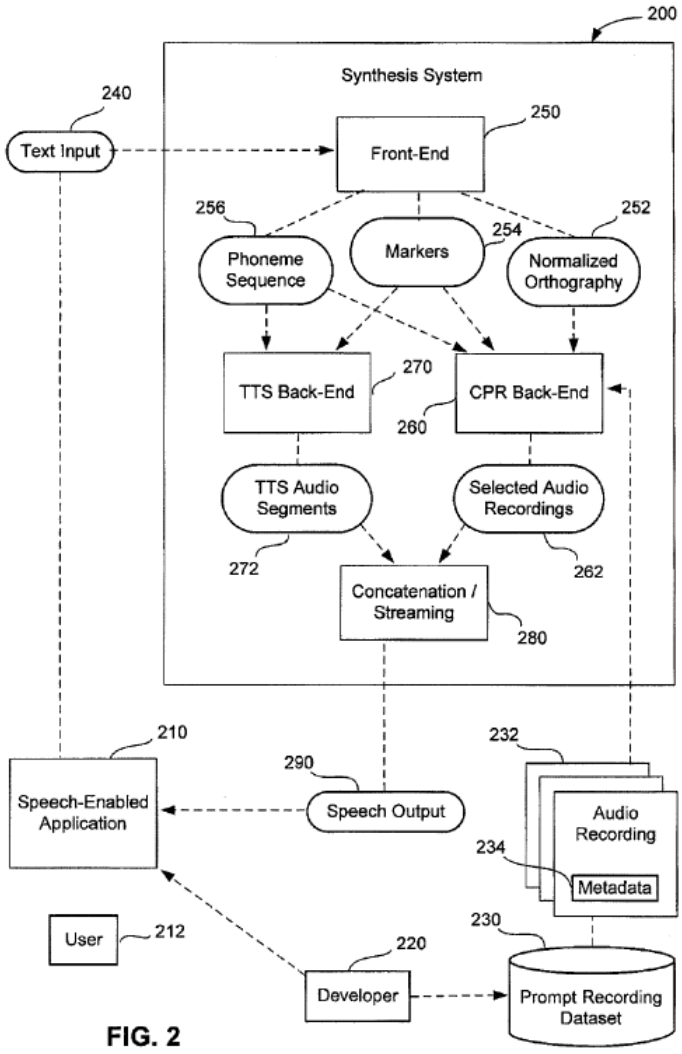


FIG. 2

Ex. 1001 (*'291Pat*), FIG. 2.

40. Using the example of figure 3A below, "the [begin stress] and [end stress] markers 340 delineate the word 'eleven' within the time '11:45 a.m.' as the specific portion of the speech output that should carry contrastive stress. In this example, '11:45 a.m.', the new time of the flight departure, contrasts with '10:45 a.m.', the original time of the flight departure. Specifically, 'eleven' is the part of

‘11:45 a.m.’ that differs from and contrasts with the ‘ten’ of ‘10:45 a.m.’ (i.e., the ‘forty_five’ and the ‘a.m.’ do not differ or contrast).” Ex. 1001 (*’291Pat*), 22:64-23:5.

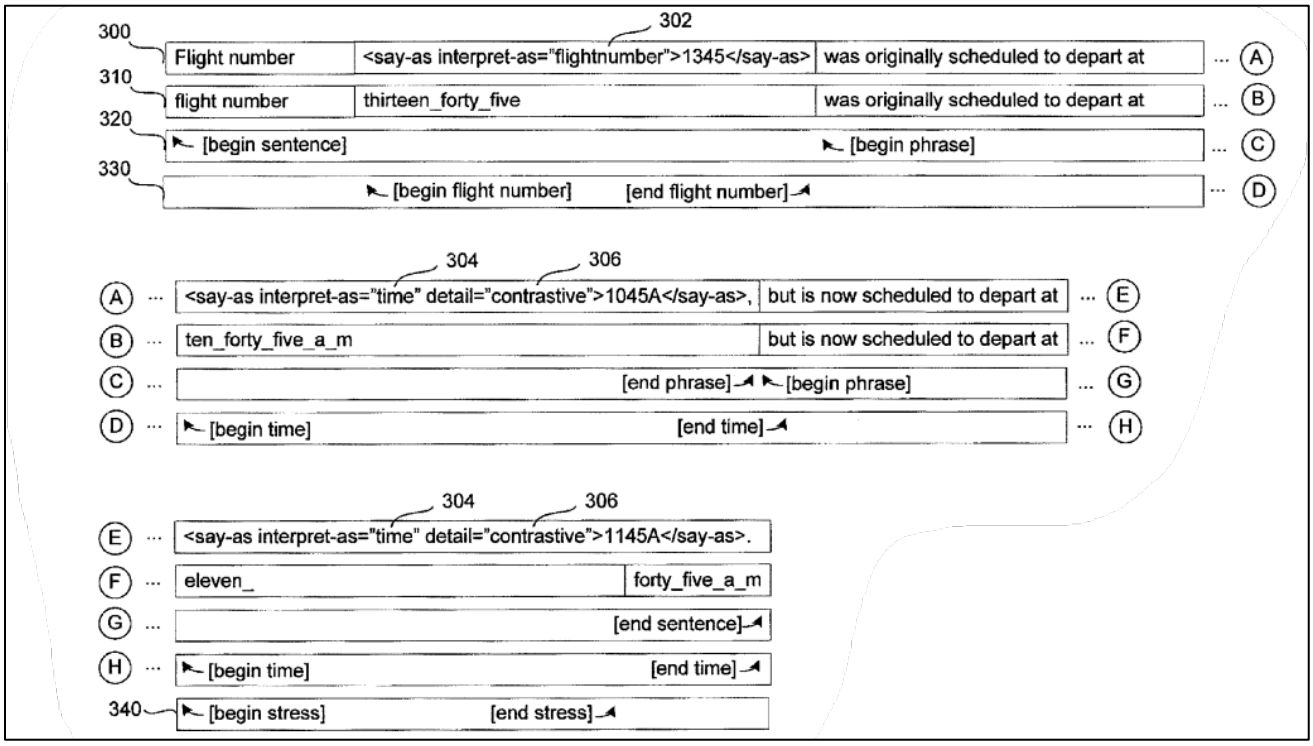
41. The ’291 patent further discloses that the identification of portions of text that should contrast may be identified through “normalized orthography” which encompasses converting text into full word forms based on the context of the text.

> Front-end 250 may also be programmed to ***expand abbreviations and acronyms into full words and/or word sequences, and to convert numerals, symbols and other meaningful characters to word forms***, using appropriate language-specific rules based on the context in which these items occur in text input 240. Numerous other examples of processing steps that may be incorporated in ***generating a normalized orthography*** 252 are possible, as the examples provided above are not exhaustive. ***Techniques for normalizing text are known***, and aspects of the present invention are not limited to any particular normalization technique.

Ex. 1001 (*’291Pat*), 15:3-14.

42. Using the example of figure 3A, “10:45 a.m.” and “11:45 a.m.” are both flight times, and therefore, “are of the same text normalization type (e.g., ‘time’) … .” *Id*., 24:2-15. Identification of the text normalization type aids

synthesis system 200 in "identifying portions of a text input that are meant to contrast with each other … ." *Id*., 24:16-20; *See also id*., 16:20-42.



Ex. 1001 (*'291Pat*), FIG. 3A.

VIII. OVERVIEW OF THE PRIMARY PRIOR ART REFERENCES

A. Analogous Art

43. I have analyzed the prior art references used in the grounds below, and it is my opinion that they are analogous to the '291 patent because they are within the same field of endeavor and reasonably pertinent to at least one problem addressed by the '291 patent.

44. In particular, *Walker* discloses a text-to-speech method and system that converts written text into audible speech. *See* Ex. 1005 (*Walker*), ¶[0002]. Like the '291 patent, *Walker* is concerned with the naturalness and quality of synthesized speech. *See id*., ¶¶[0019], [0025], [0028]. *Walker* discloses preprocessing text to determine the context of the text (i.e., context detection) and modify[ing] the text based on the context of the text (i.e., text cleaning). *See id*., ¶[0025]. After this preprocessing, cleaned text is provided to a TTS engine to synthesize text into speech which is then provided to a user. *See id*., ¶[0034].

45. *Matsumoto* discloses a speech synthesis system that converts written text into audible speech. *See* Ex. 1006 (*Matsumoto*), ¶¶[0017]-[0019], [0036]-[0037]. Like the '291 patent, *Matsumoto* is concerned with providing a user natural and easy-to-understand speech. *See id*., ¶¶[0015], [0122]. *Matsumoto* discloses taking input text, analyzing input text in order to determine how to adjust the prosody of words, and outputting synthesized speech. *See id*., ¶¶[0036]-[0043].

46. *Malsheen* discloses a text-to-speech converter for improving the handling of text that includes "abbreviations, acronyms, initialisms, non-standard compound words, numbers, dates, telephone numbers, fractions, times, and the like." Ex. 1007 (*Malsheen*), 1:5-10. Like the '671 patent, *Malsheen* is concerned with improving speech synthesizers. *See id*., 2:51-59. Specifically, *Malsheen*

discloses a system for "systematically pre-processing text that would not be properly converted into spoken words by a conventional TTS converter" (*id*.) by employing procedures to properly classify and expand text (*see id*., Abstract).

47. *Bellegarda* discloses a speech synthesis system that aims to "generat[e] variations in synthesized speech to produce speech that sounds more natural." Ex. 1014 (*Bellegarda*), 1:9-11. Like the '671 patent, *Bellegarda* is concerned with improving the intelligibility of synthesized speech. *See id*., 1:40-53. *Bellegarda*'s speech synthesis system uses a word prominence specification system that applies word prominence to newly synthesized words that are different from previously synthesized words. *See id*., 5:8-11, 5:21-26, 5:60-67.

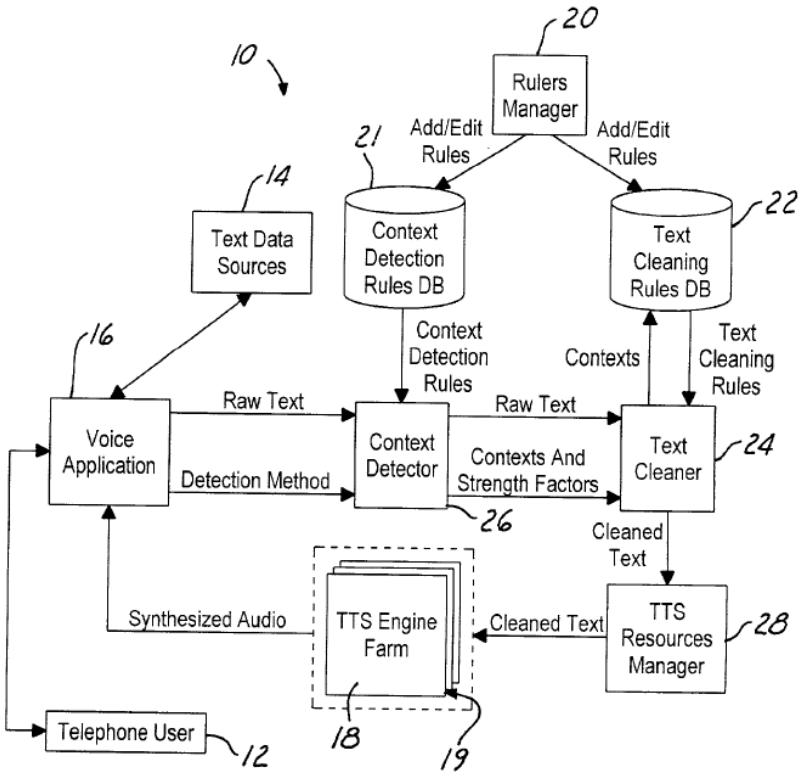
B. *Walker*

48. *Walker* relates to "a text-to-speech method and system which convert[s] written text into audible speech as a function of the context of the text." Ex. 1005 (*Walker*), ¶[0002]. *Walker* explains that while "TTS engines do their best to synthesize the written words of a text document into computer generated speech understandable by humans … the result is often an unnatural speech delivery because of the diversity and context of written words." *Id*., ¶[0003].

49. *Walker* further explains that "TTS engines are general purpose tools which deal with text processing in a general way and do not perform satisfactorily

when presented with out of the ordinary text." *Id*., ¶[0004]; *see also id.*, ¶[0005]. Thus, "without knowing the context of the document, the TTS engine may incorrectly convert [] text [] into speech." *Id*., ¶[0004]. In combination with these issues, *Walker* discloses a communication system that "convert[s] written text into audible speech as a function of the context of the text" (*id*., ¶[0006]) to "enhance the quality of a TTS conversion of the text into speech" (*id*., ¶[0007]).

50. As shown in figure 1, *Walker* discloses "communication system 10" which is "a voice portal platform for enabling a telephone user 12 to access written text … from text data sources 14." *Id*., ¶[0023], FIG. 1. "[T]ext data sources 14 may include text data sources such as the Internet and text data source providers such as email providers, news providers, ***weather condition providers***, sport scores providers, stock quotes providers, and other text data storage networks." *Id*., ¶[0023].



Ex. 1005 (*Walker*), FIG. 1.

51. *Walker* discloses that "[i]n response to a request for text or information from telephone user 12, communication system 10 locates and converts the requested text into speech and then provides the speech to the telephone user via a voice application 16." *Id*., ¶[0023].

52. For example, *Walker* discloses that a "telephone user 12 may request a weather report for a particular city" and that in response to this request, "voice application 16 accesses text data sources 14 to find a text document having the weather report for the particular city" and then "receives an electronic copy of the weather report text from a text data source 14." *Id*., ¶[0024].

53. *Walker* further discloses preprocessing components "to determine the context of the text, i.e., context detection, and then modify[ing] the text based on the context of the text, i.e., text cleaning." *Id*., ¶[0025].

54. Specifically, *Walker*'s preprocessing includes a context detector 26 that can use "context detection rules to search the text for key words and phrases associated with each context detection rule set in order to determine the context of the text" (*id*., ¶[0027]), and a text cleaner (or text normalizer) 24 that can modify "specific words, phrases, abbreviations, acronyms, and pronunciation in the text in accordance with a text cleaning rule set to modify the text so that the modified text sounds natural when converted into speech" (*id*., ¶[0028]).

55. After this preprocessing, cleaned text is provided to a "TTS resource manager 28" that "distributes the cleaned text to the appropriate TTS engine 18" in TTS engine farm 19 to synthesize text into speech. *Id*., ¶[0034].

56. For example, *Walker* discloses that "a TTS engine 18 of a TTS engine farm 19 in communication system 10 converts or synthesizes the weather report text from voice application 16 into computer generated audio speech" and then "provides the audio speech of the weather report text to telephone user 12 via voice application 16." *Id*., ¶[0024].

C. *Matsumoto*

57. *Matsumoto* relates to a "technology capable of providing a hearer with [] easy-to-hear synthetic speech." Ex. 1006 (*Matsumoto*), Abstract. *Matsumoto* explains that in certain instances such as a "weather forecast," speech synthesizers receive "almost the same sentences" and provides the following example:

> [Today's weather in the Hokkaido region is fair.] ([kyou no Hokkaido chihou no tenki wa hare desu.]), [Today's weather in the Tohoku region is fair.] ([kyou no Tohoku chihou no tenki wa hare desu.]), [Today's weather in the Kanto region is cloudy.] ([kyou no Kanto chihou no tenki wa kumori desu.]), … [Today's weather in the Kyushu region is cloudy.] ([kyou no Kyushu chihou no tenki wa kumori desu.]).

Id., ¶[0005].

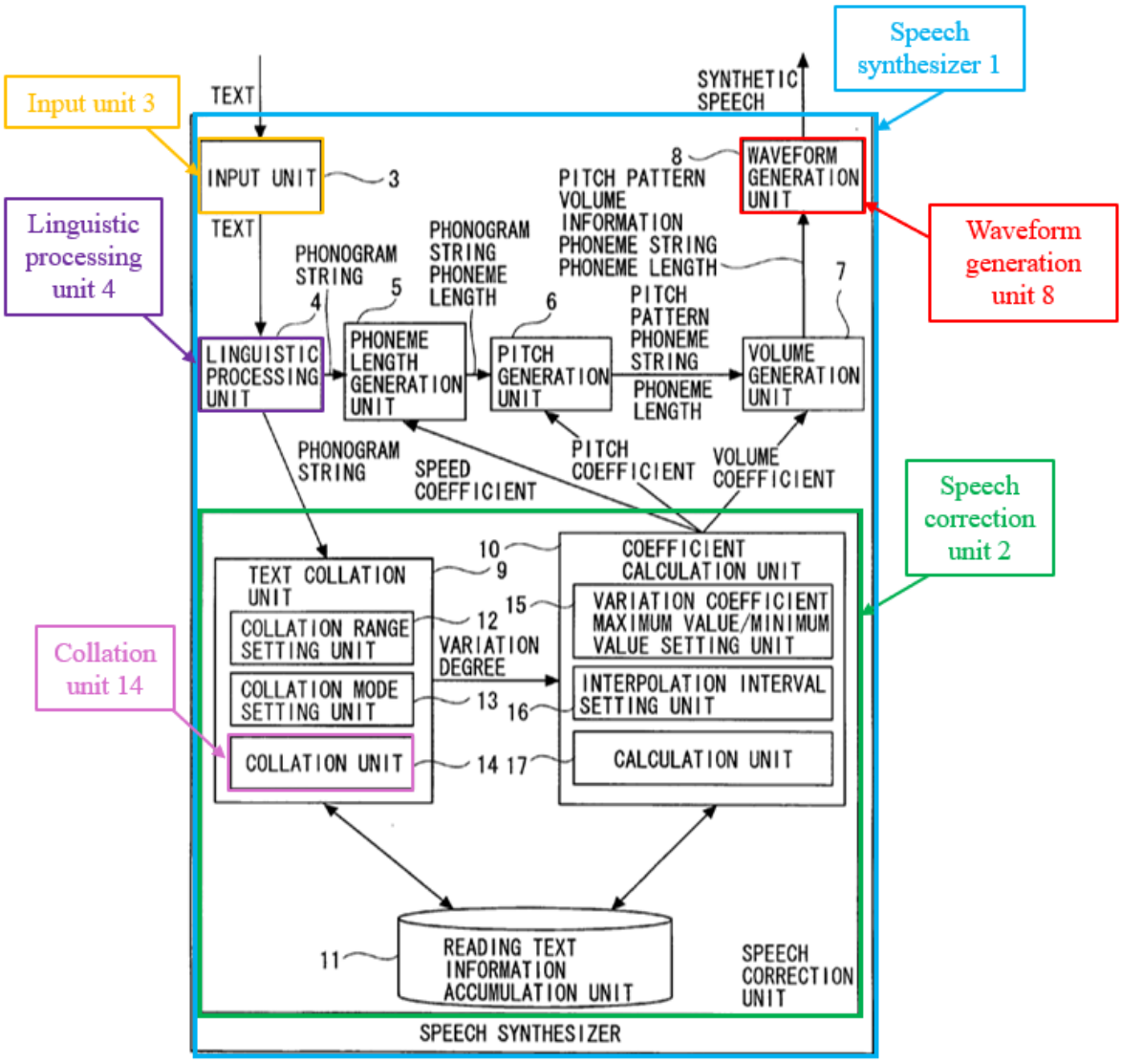
58. However, if "the speech synthesizer vocalizes such sentences in a monotone … the hearer can not [sic] concentrate on a want-to-hear point in the speech and might fail to hear the want-to-hear point." *Id*.

59. Accordingly, *Matsumoto*'s objective is to "provide a technology capable of providing the hearer with the easy-to-hear synthetic speech" (*id*., ¶[0015]), that "exhibit[s] high naturalness" (*id*., ¶[0122]) by controlling "prosodic change[s] in the sentence" (*id*.).

60. A person of ordinary skill in the art would have understood that prosodic changes, as described in *Matsumoto*, include variations in speed, pitch,

and volume when words/phrases are synthesized into speech. *See id.*, ¶¶[0004], [0040]-[0042], [0103]-[0105].

61. As shown in figure 1, *Matsumoto* discloses a "speech synthesizer 1" that "includes a speech correction unit 2, an input unit 3, a linguistic processing unit 4 … and a waveform generation unit 8." *Id.*, ¶[0036].



Ex. 1006 (*Matsumoto*), FIG. 1 (annotated).

62. "[I]nput unit 3 accepts text data of a sentence" (*id*., ¶[0037]) and "linguistic processing unit 4 executes a morphological analysis about the text (sentence) and segments this text (sentence) into morphemes (the minimum unit having a meaning in language)" (*id*., ¶[0038]).

63. After the text is segmented, "speech correction unit 2 accumulates the phonogram strings (sentences) acquired from the text data inputted by the input unit 3, then obtains a variation degree" between the accumulated sentences. *Id*., ¶[0044]. The variation degree is calculated by collation unit 14 (*see id*., ¶¶[0048],[0077]) and describes how each word or phrase in a sentence differs from a corresponding word or phrase in a previous sentence. *See id*., ¶¶[0044], [0074].

64. In particular, *Matsumoto* discloses that each sentence is assigned a sequential value "n" (with n=1 being the oldest sentence) and each word or phrase is assigned a sequential value "b":

> For instance, t(1) represents the time when the sentence of [Today's weather in the Tohoku region is fair.] ([kyou no tohoku chihou no tenki wa hare desu]) is inputted or accumulated.

Id., ¶[0072].

> A variable b is a numeral specifying, in the case of segmenting each sentence to be collated into a plurality of parts, a position of each part. Each sentence to be collated is segmented into the plurality of parts according to the same predetermined rule. For example, in the present embodiment, the

sentence is segmented into the plurality of phrases (parts) through the morphological analysis. In the example shown in FIG. 6, each of five sentences is segmented into six phrases (parts). In FIG. 6, for example, "b=1" specifies words (phrases) such as [today's] ([kyou no]), [today's] ([kyou no]), [today's] ([kyou no]), [tomorrow's] ([asu no]) and [tomorrow's] ([asu no]). Further, "b=2" specifies words such as [Tohoku], [Kanto], [Tokai], [Kansai] and [Kansai].

Id., ¶[0073].

Thus, the phrase is designated by n and b. Let a(n, b) be this phrase. In this case, for example, a(1, 2) represents [Tohoku], and a(2, 2) represents [Kanto].

Id., ¶[0074]; *see also id*., ¶[0075].

65. Accordingly, each word or phrase is described by the coordinates "a(n, b)." *See id*., ¶¶[0074]-[0075], FIG. 6.

	b=1	b=2	b=3	b=4	b=5	b=6
t(n=1)=1	KYOU NO	TOHOKU	CHIHOU NO	TENKI WA	HARE	DESU
t(n=2)=2	KYOU NO	KANTO	CHIHOU NO	TENKI WA	HARE	DESU
t(n=3)=3	KYOU NO	TOKAI	CHIHOU NO	TENKI WA	KUMORI TOKIDOKI AME	DESU
t(n=4)=4	ASU NO	KANSAI	CHIHOU NO	TENKI WA	HARE	DESU
t(n=5)=5	ASU NO	KANSAI	CHIHOU NO	SAITEIKION WA	JUDO	DESU

Ex. 1006 (*Matsumoto*), FIG. 6.

66. The collation unit 14 then "compares … two sets of a(n, b) having the same value of the variable b and different values of the variable n." *Id*., ¶[0074].

For example, as shown above, "in the process of the collation between a(1, 1) ([today's]) ([kyou no]) and a(2, 1) ([today's]) ([kyou no]), [the collation unit 14] judges that contents of the phrases are the same [and] in the collation between a(1, 2) ([Tohoku]) and a(2, 2) ([Kanto]), judges that the contents of the phrases are different." *Id*.

67. The calculated variation degree obtained by collation unit 14 is used to "calculate[] the variation coefficients (the speed coefficient, the pitch coefficient and the volume coefficient)" between each corresponding word or phrase. *Id*., ¶[0054]. The variation coefficients are then used to adjust the speed, pitch, and volume for each word or phrase. *See id*., ¶¶[0103]-[0105].

68. *Matsumoto* further discloses that "waveform generation unit 8" generates synthetic speech "from the phoneme string, the phoneme length, the pitch pattern and the volume information" based on the variation coefficients and outputs synthetic speech to "the voice output device (not shown) such as the speaker connected to the speech synthesizer 1." *Id*., ¶[0107].

IX. CLAIM CONSTRUCTION

69. I have given all the claim terms their plan and ordinary meaning, as would have been understood by a person having ordinary skill in the art at the time of the alleged invention, having taken into consideration the language of the

claims, the specification, and the prosecution history of record. In my opinion, a person of ordinary skill in the art's understanding of the plain and ordinary meaning of the claims would be the same whether the date of the alleged invention is February 12, 2010 or August 9, 2010.

X. SPECIFIC GROUNDS FOR CHALLENGE

70. In the sections below, I explain in detail how claims 1-4, 6-11, 13-17, and 19-20 of the '291 patent are not patentable over the prior art. More particularly:

71. **Ground 1A** challenges claims 1-4, 6-11, 13-17, and 19-20 as obvious over the combination of *Walker*, *Matsumoto*, and *Malsheen.*

72. **Ground 1B** challenges claims 1-4, 6-11, 13-17, and 19-20 as obvious over the combination of *Walker*, *Matsumoto*, *Malsheen*, and *Bellegarda*.[4]

[4] To the extent my declaration refers to prior art references other than those recited specifically in each of the above-listed grounds 1A and 1B, I am using such references as evidence supporting what I believe would have been known to a person of ordinary skill in the art at the time of the alleged invention, for example, about the state of the art and/or the benefits and advantages of certain design choices.

XI. DETAILED EXPLANATION OF GROUNDS

A. Ground 1A: *Walker* in Combination With *Matsumoto* And *Malsheen* Renders Obvious Claims 1-4, 6-11, 13-17, and 19-20

1. Independent Claim 1

a. 1[pre]: A method for providing speech output for a speech-enabled application, the method comprising:

73. I understand that "[a] method for providing speech output for a speech-enabled application, the method comprising" is the preamble to claim 1 of the '291 patent. I have been asked to assume that the preamble is a claim limitation. Under that assumption, in my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches the preamble.[5]

74. *Walker* discloses a method in which "[i]n response to a request for text or information from telephone user 12, communication system 10 locates and converts the requested text into speech and then provides the speech to the telephone user via a voice application 16." Ex. 1005 (*Walker*), ¶[0023], FIG. 1. More particularly, a voice application converts the audio request from the telephone user "into a text request in order to retrieve a text document related to the information requested by the telephone user." *Id*., ¶[0018].

[5] I have been informed by counsel that that the term "teaches" includes both express teaching as well as those fairly suggested to a person of ordinary skill in the art.

75. An example of a text to be converted to speech provided in *Walker* is a weather report.

> Communication system 10 is a voice portal platform for enabling a telephone user 12 to access ***written text such as*** email, news, ***weather conditions***, sport scores, stock quotes, and other information from text data sources 14.

Id., ¶[0023].

76. A TTS engine 18 within *Walker*'s system converts the text into synthesized speech and provides the audio speech of the text weather report to telephone user 12 via voice application 16:

> In response to a request for text or information from telephone user 12, communication system 10 locates and converts the requested text into speech and then provides the speech to the telephone user via a voice application 16 … ***For example, telephone user 12 may request a weather report for a particular city***. In response to this request, voice application 16 accesses text data sources 14 to find a text document having the weather report for the particular city. Voice application 16 then receives an electronic copy of the weather report text from a text data source 14. As will be described in greater detail below, a ***TTS engine 18 of a TTS engine farm 19 in communication system 10 converts or synthesizes the weather report text from voice application 16 into computer generated audio speech***. TTS engine 18 then provides the audio speech of the weather report text to telephone user 12 via voice application 16.

Id., ¶¶[0023]-[0024]; *see also id*., ¶[0034] ("Voice application 16 then forwards or streams the speech to telephone user 12 in order to satisfy the information request from the telephone user.").

77. *Walker*'s voice application 16 is consistent with how the '291 patent describes a "speech enabled application":

> Speech-enabled software applications exist that are capable of providing output to a human user in the form of speech.

Ex. 1001 (*'291Pat*), 1:26-28.

> Speech-enabled applications are used in many different contexts, such as telephone call centers for airline flight information, banking information and the like, global positioning system (GPS) devices for driving directions, e-mail, text messaging and web browsing applications, handheld device command and control, and many others …

Id., 1:31-53.

78. The '291 further describes various exemplary speech-enabled applications, including interactive voice response ("IVR"):

> One illustrative application for the techniques described herein is for use in connection with an interactive voice response (IVR) application, for which speech may be a primary mode of input and output … Examples of speech-enabled applications include, but are not limited to, automated call-center applications, internet-based applications, device-based applications, and any other suitable application that is speech enabled.

Id., 9:35-60.

79. Accordingly, a person of ordinary skill in the art would have understood *Walker* to teach a system having a voice application for performing a method of converting text into speech ("a method for providing speech output for a speech-enabled application").

b. 1[a]: receiving from the speech-enabled application a text input comprising a text transcription of a desired speech output;

80. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

81. *Walker* discloses that context detector 26 receives raw text (i.e., unprocessed text) from voice application 16. Ex. 1005 (*Walker*), ¶[0030] (describing the "electronic copy of the text document provided by voice application 16 to context detector 26 is labeled 'Raw Text' in FIG. 1"), FIG. 1. As I discussed in §XI.A.1.a (limitation 1[pre]), a person of ordinary skill in the art would have understood that voice application 16 is a speech-enabled application. Thus, a person of ordinary skill in the art would have understood context detector 26 receives raw text from a speech-enabled application 16.

82. *Walker* further discloses that one type of text received by the context detector 26 from voice application 16 is "a weather report for a particular city." *Id*.,

¶[0024]. While *Walker* discloses that the weather report is for a particular city, in my opinion, it would have been obvious to a person of ordinary skill in the art that *Walker*'s weather report could pertain to more than one location/city or day. Indeed, even for a single location, weather reports usually describe the weather for multiple days, e.g. "today" and "tomorrow." For example, *Matsumoto* discloses a weather forecast to include the weather for different days (i.e., today "kyou no" and tomorrow "asu no"):

> In FIG. 6, a variable n corresponds to a numeral for designating each sentence. For example, "n=1" corresponds to the numeral for specifying a sentence of [***Today's*** weather in the Tohoku region is fair.] ([***kyou no*** tohoku chihou no tenki wa hare desu]), and "n=2" corresponds to the numeral for specifying a sentence of [***Today's*** weather in the Kanto region is fair.] ([***kyou no*** kanto chihou no tenki wa hare desu]). "n=5" corresponds to a sentence of [The ***tomorrow's*** lowest temperature in the Kansai region is 10 degrees.] ([***asu no*** kansai chihou no saitei kionn wa juudo desu.]), and this sentence, in the example in FIG. 6, is shown as a sentence inputted afresh to the speech correction unit 2 (the speech synthesizer 1).

Ex. 1006 (*Matsumoto*), ¶[0071]; *see also id*., ¶¶[0005], [0073].

83. Additionally, *Matsumoto* discloses a weather forecast to include the weather for different locations. *See id*., ¶[0071] (disclosing weather for the "Tohoku region," the "Kanto region," and the "Kansai region.").

84. In my opinion, a person of ordinary skill in the art would have understood weather reports and weather forecasts to include similar information.

85. It is my opinion that it would have been obvious to modify *Walker* based on *Matsumoto*'s disclosure such that the context detector 26 receives from the voice application 16 raw text that pertains to the weather report for more than one location/city because, as *Matsumoto* teaches, a user (e.g., a telephone user 12 interacting with voice application 16 in *Walker*'s system) may be interested in the weather from more than one location/city.

86. In my opinion, the combination would also have been obvious to a person of ordinary skill in the art because it would have amounted to combining prior art elements (*Walker* teaches context detector 26 receiving a text weather report) according to a known method (*Matsumoto* teaches that a text weather report can pertain to more than one location/city or day) to yield predictable results (the context detector receiving a text weather report that pertains to more than one location/city or day).

87. It is also my opinion that the combination would have been obvious to a person of ordinary skill in the art because it amounted to a simple substitution of a known element (context detector receiving a text weather report for a particular city) for another (context detector receiving a text weather report for more than one

location/city or day) to obtain predictable results (to determine the context of text as pertaining to a weather report).

88. Thus, in my opinion, a person of ordinary skill in the art would have found it obvious for *Walker*'s weather report to include two or more sentences (or text input), each describing the weather for a different location/city or day, as taught by *Matsumoto*.

89. As I discussed in §VIII.B (overview of *Walker*), *Walker* converts written text into synthesized speech output. In my opinion, a person of ordinary skill in the art would have understood the synthesized speech output of the *Walker-Matsumoto* combination to be a "desired speech output" because *Walker* discloses a user using *Walker*'s system to request that a text weather report be outputted as synthesized speech:

> ***In response to a request for text or information from telephone user*** 12, communication system 10 locates and converts the requested text into speech and then provides the speech to the telephone user via a voice application 16.

Ex. 1005 (*Walker*), ¶[0023].

> For example, ***telephone user 12 may request a weather report for a particular city***. In response to this request, voice application 16 accesses text data sources 14 to find a text document having the weather report for the particular city. Voice application 16 then receives an electronic copy of the

> weather report text from a text data source 14. As will be described in greater detail below, a TTS engine 18 of a TTS engine farm 19 in communication system 10 converts or synthesizes the weather report text from voice application 16 into computer generated audio speech. TTS engine 18 then provides the audio speech of the weather report text to telephone user 12 via voice application 16.

Id., ¶[0024].

90. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach a context detector receiving from a voice application ("receiving from the speech-enabled application") the raw text of a weather report that a user wanted to hear as speech output ("a text input comprising a text transcription of a desired speech output").

c. 1[b]: identifying at least one first portion of at least one token of the text input that differs from at least one corresponding first portion of at least one other token of the text input, and at least one second portion of the at least one token that does not differ from at least one corresponding second portion of the at least one other token;

91. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

92. In the discussion that follows, I first discuss a person of skill in the art's understanding of "token" as used in the claims and specification of the '291 patent. Next I discuss why a POSITA would have found it obvious to combine the

teachings of *Walker* and *Matsumoto*. Then, I discuss an example text weather report that a person of skill in the art would have understood to be taught by the *Walker-Matsumoto* combination. Thereafter, I discuss why a person of skill in the art would have found it obvious to further modify the *Walker-Matsumoto* combination based on *Malsheen* to further process the example text weather report prior to converting it into synthesized audio output. Finally, I explain how the *Walker-Matsumoto-Malsheen* combination would operate on the example text weather report.

(i) "Token" as used in the '291 claims and specification

93. In my opinion, a person of skill in the art would have understood "token" as used in the claims and specification of the '291 patent to encompass a word or lexical unit of text separated by white spaces:

> In some embodiments, the synthesis system may automatically identify which ***tokens (e.g., words)*** should contrast with each other without any tags being included in the text input.

Ex. 1001 (*'291Pat*), 9:13-15.

> Synthesis system 200 may be programmed, e.g., through a ***tokenizer*** implemented as part of front-end 250, to parse text input 350 into a sequence of ***individual tokens on the order of single words***. In the illustration of FIG. 3B, the ***individual parsed tokens are represented as separated by white space*** in the normalized orthography 360.

Id., 31:56-61; *see also id*., 25:62-67 ("words/word sequences (or equivalently, tokens/token sequences)"), 26:2-5 ("one or more linking tokens such as 'originally', 'but' and 'is now'"), 26:19-20 ("linking token 'originally'"); 26:21 ("linking tokens 'but is now'"), 36:23-25 ("At act 410, the text input may be tokenized, i.e., parsed into individual tokens on the order of single words.").

94. Indeed, this is consistent with how a person of skill in the art would have understood the term "token" to be used in the art.

95. For example, W3C Speech Synthesis Markup Language (SSML) Ver. 1.1 WD4 , September 4, 2007 ("*W3C*"), developed by the Voice Browser Working Group, is a well-known specification that provides a standard for implementing the SSML mark-up language and generating synthetic speech. *W3C* discusses "a standard way to control aspects of speech such as pronunciation, volume, pitch, rate, etc. across different synthesis-capable platforms." Ex. 1008 (*W3C*), at 1. *W3C* also describes tokenization:

> In English, ***tokens are usually separated by white space and are typically words***. For languages with different tokenization behavior, the term "word" in this specification is intended to mean ***an appropriately comparable unit***.

Id., 5.

96. In my opinion, a person of skill in the art would have also understood the '291 patent to use the terms "token" and "field" interchangeably:

> In some embodiments, the application may include tags in the text input to identify ***tokens or fields*** that should contract with each other[.]

Ex. 1001 (*'291Pat*), 9:8-10.

> Identifying specific portions of one or more ***fields or tokens*** to be rendered to carry contrastive stress …

Id., 28:62-63.

> In some embodiments, the entire ***token "1145A"*** may be identified to carry contrastive stress, while in other embodiments synthesis system 200 may next proceed to identify only one or more specific portions of the ***field "1145A"*** to be rendered to carry contrastive stress.

Id., 27:56-60; *see also id.*, 32:63-66 (referring to "10:30 a.m.," "11:30 a.m.," and "12:30 p.m." as tokens), 33:16-21 (referring to "10:30 a.m.," "11:30 a.m.," and "12:30 p.m." as fields).

97. The '291 patent explains that tokens may have a text normalization type, and that a normalized orthography corresponding to the inputted text token may be generated. *See* Ex. 1001 (*'291Pat*), 36:29-36 ("As discussed above, text normalization types of various tokens may be identified … [and] a normalized orthography corresponding to the text input may be generated.").

98. Examples of text normalization types mentioned in the '291 patent include state names. *See id*., 24:45-51. The '291 patent describes "normalized

orthography" as encompassing the conversion of abbreviations into full words or word sequences.

> Front-end 250 may also be programmed to expand ***abbreviations*** and acronyms ***into full words and/or word sequences***, and to convert numerals, symbols and other meaningful characters to word forms, using appropriate language-specific rules based on the context in which these items occur in text input 240. Numerous other examples of processing steps that may be incorporated in generating a normalized orthography 252 are possible, as the examples provided above are not exhaustive. Techniques for normalizing text are known, and aspects of the present invention are not limited to any particular normalization technique.

Id., 15:3-14.

> [T]he normalized orthography may represent a standardized spelling out of the words included in the text input, which for some tokens may depend on their text normalization type.

Id., 36:37-40.

(ii) Motivation to Combine *Walker* and *Matsumoto*

99. In my opinion, a person of ordinary skill in the art would have been motivated to combine the teachings of *Walker* and *Matsumoto*, and would have had a reasonable expectation of success in doing so.

100. *Walker* teaches a system and method for improving the naturalness of synthesized speech using the text's context. *See* §VIII.B (overview of *Walker*).

Walker teaches obtaining more natural and higher quality speech synthesis by preprocessing text in certain ways, for example, by using a context detector and text cleaner to process raw text received by a voice application:

> In operation, ***context detector 26*** is operable with voice application 16 to ***receive an electronic copy of the text document*** obtained from text data source 14 in response to a request for information from telephone user 12. This electronic copy of the text document provided by voice application 16 to context detector 26 is labeled 'Raw Text' in FIG. 1 as the text of the document obtained from text data source 14 has not been processed. ***Context detector 26 processes the raw text to determine the context of the text by locating in the text key words and phrases associated with each context detection rule set stored in context detection rules database 21.***

Ex. 1005 (*Walker*), ¶[0030].

> Context detector 26 then provides an indication of the context of the text to text cleaner 24.

Id., ¶[0031].

> Upon receiving the raw text and a context identifier from context detector 26, text cleaner 24 accesses text cleaning rules database 22 to access the text cleaning rules associated with a context of the text. ***Text cleaner 24 then replaces, modifies, or otherwise changes the raw text in accordance with the text cleaning rules to produce 'Cleaned Text' as shown in FIG. 1.***

Id., ¶[0033]; *see also id*., ¶¶[0011]-[0015], [0034]-[0038].

101. *Walker* discloses that one type of text that is pre-processed is "a weather report for a particular city" (*id*., ¶[0024]) which is then provided to a TTS resource manager 28 that "distributes the cleaned text to the appropriate TTS engine 18" in TTS engine farm 19 to synthesize text into speech (*id*., ¶[0034]).

102. *Walker*, however, does not provide extensive implementation details regarding the TTS engines 18 within the TTS engine farm 19, including as they would relate to TTS engines for synthesizing weather reports (or similar text documents) or how such a TTS engine would improve the naturalness of synthesized speech. *See id*., ¶[0034].

103. Thus, in my opinion, a person of ordinary skill in the art would have been motivated to find implementation details for a TTS engine capable of synthesizing text weather reports into speech with high naturalness to implement *Walker*'s teachings. In so looking, a person of ordinary skill in the art would have readily found *Matsumoto*.

104. *Matsumoto* discloses a speech synthesizer that aims to provide easy-to-hear synthesized speech that "exhibit[s] high naturalness" (Ex. 1006 (*Matsumoto*), ¶[0122]) for example, in a "weather forecast" (*id*., ¶[0121]). *see also id.*, ¶¶[0005], [0014], [0015], [0121].

105. In my opinion, it was well-known that speech synthesizers like *Matsumoto*'s were deployed in connection with speech applications, such as *Walker*'s voice application. Indeed, *Walker* itself discloses that its system uses a TTS engine (i.e., a speech synthesizer) working in conjunction with a voice application to convert text to synthesized speech:

> Voice application 16 then receives an electronic copy of the weather report text from a text data source 14. As will be described in greater detail below, a ***TTS engine 18*** of a TTS engine farm 19 in communication system 10 ***converts or synthesizes the weather report text*** from voice application 16 into computer generated audio speech. TTS engine 18 then ***provides the audio speech of the weather report text to telephone user 12 via voice application 16.***

Ex. 1005 (*Walker*), ¶[0024].

106. Furthermore, other TTS systems described in prior art patents and available to a person of ordinary skill in the art confirm that a voice application and TTS engine are used together.

107. For example, U.S. Pat. 7,010,489 ("*Lewis*") discloses a "preferred architecture for a TTS system in [a] computer 1" that includes a "speech enabled application 12." Ex. 1009 (*Lewis*), 4:51-56. *Lewis* also discloses that while "the TTS system 10, speech dictation system 11, and the speech enabled application 12 are shown as separate application programs … these various applications could, of

course, be implemented as a single, more complex applications program." *Id*., 4:56-62; *see also id*., FIG. 2.

108. Similarly, U.S. Pat. No. 7,177,807 ("*Schmid*") discloses "enabling speech recognition and speech synthesis technology." Ex. 1010 (*Schmid*), 1:15-16. *Schmid* discloses that its "[s]peech recognition system 200 includes one or more speech recognition applications 202[.]" *Id*., 5:36-38. "Applications 202 can be speech recognition and speech synthesis applications" (*id*., 5:49-50) which invoke "one or more text-to-speech engines (synthesizers) 208." (*id*., 5:39-40). "[A]pplications 202 may provide the source of audio data for speech recognition" (*id*., 5:54-55) and can also "provide a source of textual data to be synthesized" (*id*., 5:62-63). *See also id.*, FIG. 2.

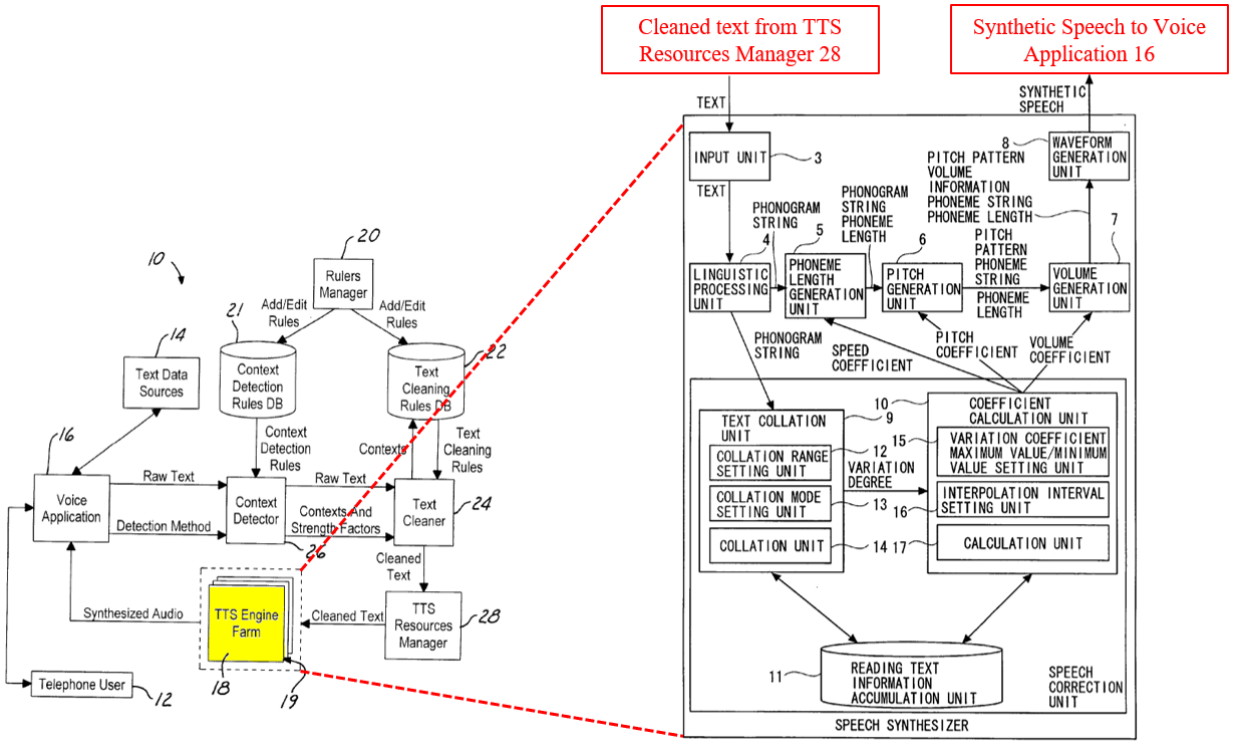
109. In my opinion, a person of ordinary skill in the art would have appreciated that *Matsumoto*'s speech synthesizer would have advanced *Walker*'s objective of providing more natural synthesized speech for a weather report and provided additional details regarding how to implement such a TTS engine.

110. As I discussed in §VIII.B (overview of *Walker*), *Walker* determines the context for a document, and then applies context rules to pre-process the text and improve the quality and naturalness of synthesized speech. Similarly, as discussed in §VIII.C (overview of *Matsumoto*) *Matsumoto* aims to provide easy-to-

hear synthesized speech with high naturalness by controlling prosodic changes (e.g., speed/duration, pitch, volume) for synthesized speech based on whether corresponding words in text are identified as different or the same.

111. In my opinion, a person of ordinary skill in the art would have known that similarities and differences between sentences of input text (e.g., a document) would cause a human speaker to emphasize some words by means of intonational features (e.g., pitch, duration, and amplitude of the sounds (phonemes)).

112. Thus, in my opinion, a person of ordinary skill in the art would have found it obvious to incorporate *Matsumoto*'s teachings regarding its speech synthesizer into *Walker* as one of *Walker*'s TTS engines 18 in TTS farm 19 for converting text weather reports into more naturally flowing, easy-to-hear synthesized speech, as conceptually illustrated in the figure below.



113. In my opinion, the combination would also have been obvious to a person of ordinary skill in the art because it would have amounted to using a known technique (*Matsumoto*'s speech synthesizer for text weather reports) in a known device (*Walker*'s text-to-speech system which converts text of weather reports into synthesized speech) ready for improvement (*Walker* discloses using available TTS engines) to yield predictable results (improving the naturalness and quality of synthesized speech of a weather report).

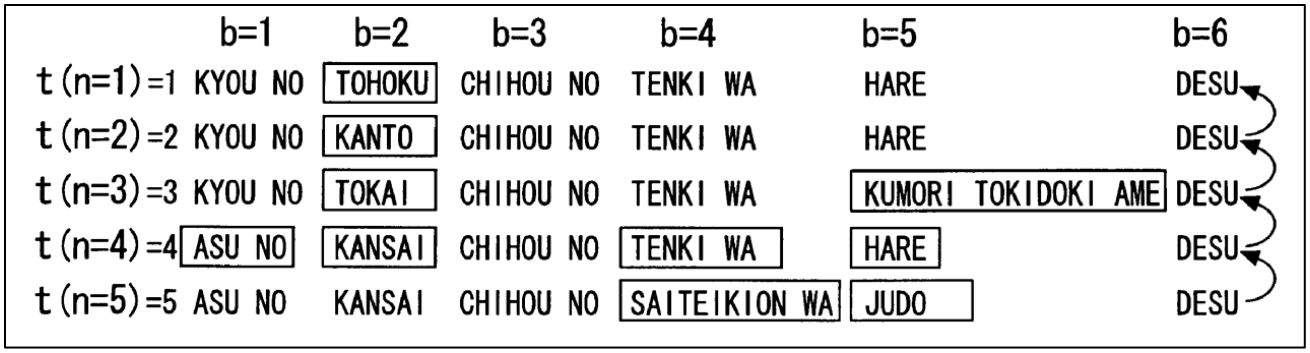
114. Further, it is my opinion that a person of ordinary skill in the art would have had a reasonable expectation of success in making the modification because *Matsumoto* discloses additional details for a TTS engine that improves the

naturalness of synthesized speech that would have been within an ordinary artisan's ability to implement in *Walker*. For example, both *Walker* and *Matsumoto* synthesize speech from documents. *See* Ex. 1005 (*Walker*), ¶¶[0004]-[0005], [0024], [0030]; Ex. 1006 (*Matsumoto*), ¶¶[0005], [0037]-[0038].

115. Additionally, *Walker* already discloses using a speech synthesizer that receives text input and outputs synthetic speech, and it would have been within a POSITA's ability to program *Matsumoto*'s speech synthesizer as one of *Walker*'s TTS engines, because *Walker* produces text which is sent to the TTS engines, and *Matsumoto* takes text as its input.

(iii) Example text weather report that a person of ordinary skill in the art would have understood to be taught by the *Walker-Matsumoto* combination

116. For the reasons discussed in §XI.A.1.b (limitation 1[a]), the *Walker-Matsumoto* combination teaches a text weather report comprising two or more sentences describing the weather for two or more locations/cities. *Matsumoto*'s figure 6 provides an example set of such sentences.



Ex. 1006 (*Matsumoto*), FIG. 6.

117. While *Matsumoto*'s examples refer to a location by only a single name (e.g., Tohoku or Kanto), a person of ordinary skill in the art would have understood that a weather report can refer to a location by city and state names, and that state names are often written using the abbreviation for that state's name, for example: The weather in Greenville, NC is fair.

118. While *Matsumoto*'s figure 6 shows five sentences (i.e., t(n=1) through t(n=5)), in my opinion, a person of ordinary skill in the art would have understood *Matsumoto*'s system to be capable of operating on two or more sentences.

119. For example, *Matsumoto* discloses its collation unit 14 comparing a given word to the word in the same "b" position in the prior sentence:

> Thus, the phrase is designated by n and b. Let a(n, b) be this phrase. In this case, for example, a(1, 2) represents [Tohoku], and a(2, 2) represents [Kanto]. ***The collation unit 14 compares, as the collation process, two sets of a(n, b) having the same value of the variable b and different values of the variable n.*** The collation unit 14, in the process of the collation between

a(1, 1) ([today's]) ([kyou no]) and a(2, 1) ([today's]) ([kyou no]), judges that contents of the phrases are the same. Moreover, the collation unit 14, in the collation between a(1, 2) ([Tohoku]) and a(2, 2) ([Kanto]), judges that the contents of the phrases are different.

Ex. 1006 (*Matsumoto*), ¶[0074].

120. *Matsumoto* discloses the collation unit 14 using Expression 1 to calculate the variation degree. *See id.*, ¶¶[0077]-[0079].

[Mathematical Expression 1]

$$v(n, b) = \sum_{m=1}^{n} \frac{1 - \delta(a(m, b), a(m-1, b))}{t(n) - t(m) + 1}$$

Ex. 1006 (*Matsumoto*), Expression 1.

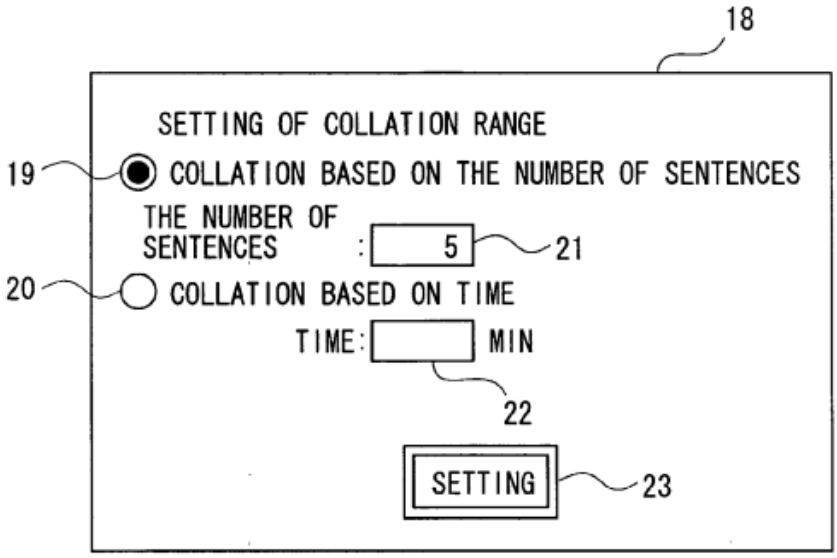
121. Additionally, *Matsumoto* discloses that the collation can be based on the number of sentences (e.g., two sentences), including comparing the latest sentence with the just-anterior sentence. *See id*., ¶¶[0058]-[0059], [0061]-[0062].

122. For example, as shown in figure 2 below, *Matsumoto* discloses that a user can set a collation range to be between a user-specified number of sentences:

> ***[T]he user chooses the selection button 19*** (a button for specifying the [collation based on the number of sentences]), ***then inputs the number of sentences*** to the sentence count input field 21, and presses the setting button 23. In this case, the collation range setting unit 12 retains the collation

method selected by the selection button 19 and the collation range (the number of sentences) inputted to the sentence count input field 21.

Id., ¶[0059].



Ex. 1006 (*Matsumoto*), FIG. 2.

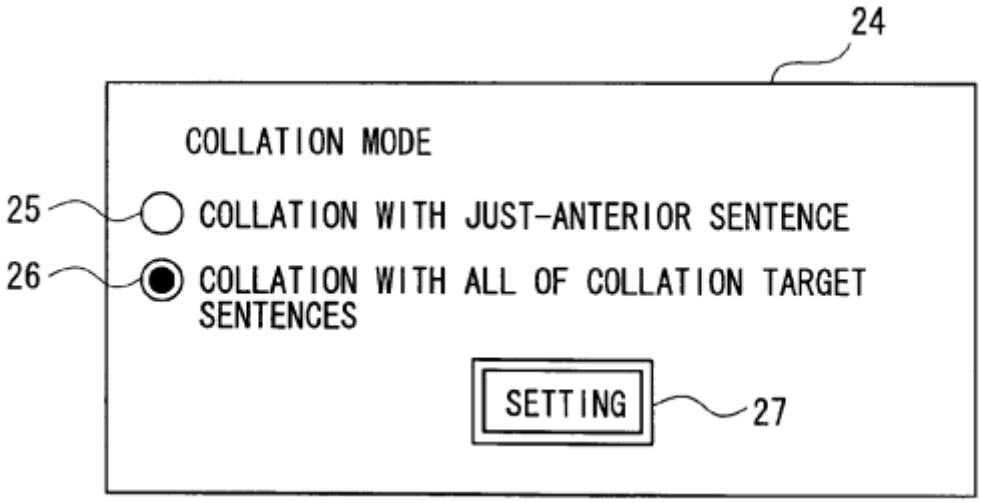
123. Indeed, *Matsumoto*'s use of radio buttons in figure 2 would have conveyed to a person of ordinary skill in the art that collation based on a number of sentences is one of two alternative choices (i.e., the other choice being a collation based on time). Thus, in my opinion, a person of ordinary skill in the art would have understood that the collation range could be based on two sentences.

124. As shown in figure 3 below, *Matsumoto* also discloses that a user can set a collation mode to be with the just-anterior sentence (i.e., the prior sentence):

It is assumed that the user chooses the selection button 25 (a button for specifying the mode of the [collation with just-anterior sentence] (the first

collation mode) as the collation mode), and selects the setting button 27. In this case, the collation mode setting unit 13 retains the selected collation mode (the first collation mode) as the collation mode executed in the speech synthesizer 1.

Id., ¶[0062].



Ex. 1006 (*Matsumoto*), FIG. 3.

125. Indeed, *Matsumoto*'s use of radio buttons in figure 3 would have conveyed to a person of ordinary skill in the art that collation with the just-anterior sentence is one of two alternative choices (i.e., the other choice being a collation based on all target sentences). Thus, in my opinion, a person of ordinary skill in the art would have understood that the collation mode could also be based on a collation between two adjacent sentences.

126. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* combination to teach producing a

synthetic speech output for a text weather report comprised of two sentences in which the location information includes an abbreviation representing a state name, such as the example weather report illustrated below:[6]

The weather in Greenville, NC is fair.

The weather in Greenville, SC is fair.

(iv) Motivation to combine *Walker-Matsumoto* with *Malsheen*

127. In my opinion, a person of ordinary skill in the art would have been motivated to further modify the *Walker-Matsumoto* combination in view of *Malsheen*'s disclosure and would have had a reasonable expectation of success in doing so.

128. As I detailed in the preceding section, a person of ordinary skill in the art would have understood that an example weather report such as the one below would have been encompassed within the *Walker-Matsumoto* combination's teachings:

[6] To be sure, this example weather report is not explicitly stated in *Matsumoto*. However, in my opinion, a person of ordinary skill in the art would have understood *Matsumoto*'s disclosure (and *Walker*'s as well), would encompass, and apply to, such a weather report.

The weather in Greenville, NC is fair.

The weather in Greenville, SC is fair.

129. For the reasons I discussed in §XI.A.1.c(i) ("token" as used in the '291 claims and specification), a person of ordinary skill in the art would have understood each word of the example weather report above—including the abbreviations "NC" and "SC"—to comprise a "token" as that term is used in the '291 patent claims and specification.

130. While *Walker* provides examples of expanding abbreviations into full words (*see, e.g.*, Ex. 1005 (*Walker*), ¶[0033]), it does not discuss in great detail how the context cleaner should modify or change abbreviations representing a location (e.g., representing a state name) before the text is sent to the appropriate TTS engine.

131. *Malsheen* discloses a text-to-speech converter "for improving the handling of text that includes ***abbreviations***, acronyms, initialisms, non-standard compound words, numbers, dates, telephone numbers, fractions, times, and the like." Ex. 1007 (*Malsheen*), 1:5-10.

132. *Malsheen* discloses a text classifier procedure that "determines and classifies the syntax aspects of each word" including making determinations as to whether a word is a member of a particular class (e.g., geographic names). *Id*.,

5:41-50; *see also* 6:20-39. The text classifier procedure "calls the Text Expander procedure 140" (*id*., 7:20-22), which "determine[es] what expansion of the word is required" (*id*., 7:10-12).

133. For example, additional text processing may be performed by the Text Expander procedure 140 which uses "expansion procedures (sometimes called rules)" (*id*., 4:14-15) to expand state abbreviations into a full words. *See id*., 8:31-32 ("the 'state' rule indicates that 'mo' should be expanded to 'Missouri'"); *see also id.*, 9:19-10:62 (explaining how the text "S F, CA" would be expanded to "San Francisco California.").

134. In my opinion, a person of ordinary skill in the art would have understood *Malsheen*'s reference to "expand[ing]" an abbreviation, such as a state name, to "a full word string" to mean that the abbreviation of a state is written in a full word form.

135. Further, it is my opinion that a person of ordinary skill in the art would have been motivated to incorporate *Malsheen*'s teaching that abbreviations representing a location should be converted to written words into the *Walker-Matsumoto* combination such that abbreviations appearing as a combination of letters in the text weather report are converted to full text by the text cleaner 24 to

ensure proper pronunciation by the speech synthesizer (i.e., "South Carolina" instead of "s c").

136. In my opinion, a person of ordinary skill in the art would have found the combination obvious because it amounted to the use of a known technique (*Malsheen*'s teaching to expand abbreviations into full words before sending text to a TTS engine) to improve a similar system (*Walker-Matsumoto* teaches converting text weather reports incorporating abbreviations representing locations to synthesized speech) in the same way (abbreviations representing locations in the *Walker-Matsumoto* text weather reports are converted to full words as taught by *Malsheen* before the weather report is sent to a TTS engine).

137. Additionally, a POSITA would have had a reasonable expectation of success in making the modification because *Malsheen* discloses additional details for text normalization implemented by *Walker*'s context cleaner that would have been within the ability of an ordinary artisan to implement.

138. Thus, in my opinion, a person of ordinary skill in the art would have understood that in the *Walker-Matsumoto-Malsheen* combination, the text cleaner would convert the abbreviation "NC" to "North Carolina" and the abbreviation "SC" to "South Carolina," such that the example text weather report sent to the TTS engine would comprise:

The weather in Greenville, North Carolina is fair.

The weather in Greenville, South Carolina is fair.

139. In my opinion, a person of ordinary skill in the art would have further understood that this process of expanding abbreviations of state names to full words is analogous to what the ’291 patent refers to as “normalized orthography” as I discussed in §XI.A.1.c(i) (“token” as used in the ’291 claims and specification).

(v) Application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report

140. In the discussion that follows, I assume that the weather report received by *Walker*’s voice application 16 from text data source 14 is the example weather report I provided in §XI.A.1.c(iii), and that *Matsumoto*’s TTS system is configured to perform collation based on the number of sentences (i.e., two sentences) and to collate the latest sentence with only the anterior sentence.

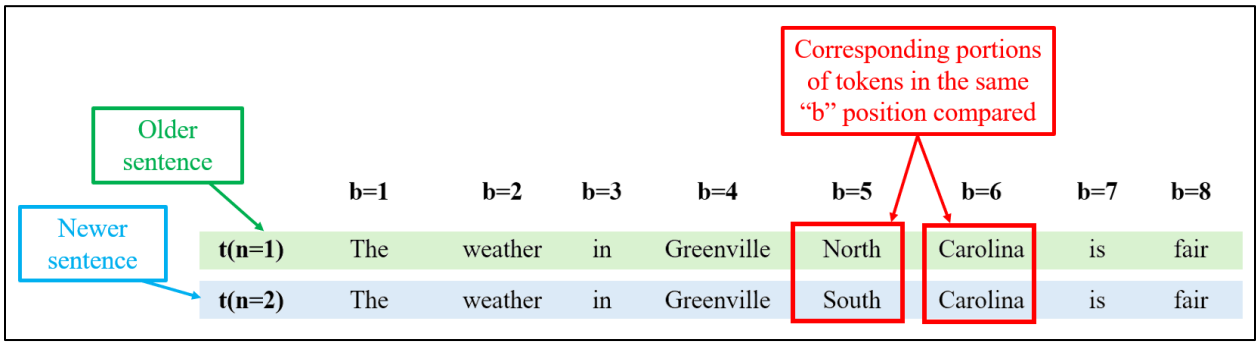
141. As I discussed in §XI.A.1.c(ii) (motivation to combine *Walker* and *Matsumoto*), a person of ordinary skill in the art would have found it obvious to incorporate *Matsumoto*’s TTS system into *Walker*’s TTS engine farm as the TTS engine for synthesizing text weather reports into speech.

142. Further, a person of ordinary skill in the art would have understood that *Walker-Matsumoto-Malsheen*’s speech synthesizer would operate on the

"cleaned" sentences of the example weather report, as I discussed in §XI.A.1.c(iv) (motivation to combine *Walker-Matsumoto* with *Malsheen*).

143. *Matsumoto* discloses its speech synthesizer includes collation unit 14 that calculates the variation degree of the new sentence from the previous sentence which describes how each word or phrase in a new sentence differs from a corresponding word or phrase in a previous sentence. Ex. 1006 (*Matsumoto*), ¶¶[0044], [0074].

144. Specifically, *Matsumoto* discloses that collation unit 14 compares two sets of a(n, b). *See id.*, ¶[0074] (disclosing that a(1, 1) "kyou no" and a(2, 1) "kyou no" are compared while a(1, 2) "Tohoku" and a(2, 2) "Kanto" are compared); *see also id.*, ¶[0075]. Thus, using the example weather report discussed herein, a(1, 5) "North" and a(2, 5) ("South") are compared while a(1, 6) "Carolina" and a(2, 6) "Carolina" are compared.



145. As shown, *Matsumoto*'s collation unit would identify, for example, that "North" is different from "South" and that "Carolina" is the same as

"Carolina," each of which constitute a "portion" of different tokens, "NC" and "SC," respectively. *See* Ex. 1006 (*Matsumoto*), ¶[0074] (discussing with reference to figure 6 that "in the process of the collation between a(1, 1) ([today's]) ([kyou no]) and a(2, 1) ([today's]) ([kyou no])," collation unit 14 "judges that contents of the phrases are the same" and "in the collation between a(1, 2) ([Tohoku]) and a(2, 2) ([Kanto]), judges that the contents of the phrases are different.").

146. *Matsumoto* further discloses that when corresponding words or phrases between two sentences are determined to be the same (e.g., "Carolina" and "Carolina") the variation degree is "0," but if determined to be different (e.g., "North" and "South"), the variation degree is "1." *See id*., ¶[0080]; Expression 1.

147. For example, below I use the example of "South" and Expression 1 to explain how *Matsumoto* calculates a variation degree of 1 (i.e., when words are determined to be different).

148. First, a person of ordinary skill in the art would have understood that in the summation of Expression 1, $\sum_{m=1}^{n}$, variable "m" ranges from 1 to 2 because two sentences (i.e., n=1 and n=2) are collated. *See id.*, ¶¶[0079]-[0080].

[Mathematical Expression 1]

$$v(n, b) = \sum_{m=1}^{n} \frac{1 - \delta(a(m, b), a(m-1, b))}{t(n) - t(m) + 1}$$

Ex. 1006 (*Matsumoto*), Expression 1.

149. When m=1, the expression δ(a(m, b), a(m−1, b)) is evaluated as δ(a(1, b), a(1−1, b)), i.e., δ(a(1, b), a(0, b)). As *Matsumoto* notes, "a(0, b) = a(1, b)." *Id*., ¶[0080]. Thus, when m=1, δ(a(m, b), a(m−1, b)) in Expression 1 is "1." *See id*., ¶[0080] ("[I]n the mathematical expression (1), δ(a(m, b), a(m−1, b)) represents '1' when a(m, b) is equal to a(m−1, b) … "). In this case, the first sentence (n=1) is being compared to itself.

150. When m=2, the expression δ(a(m, b), a(m−1, b)) is evaluated as δ(a(2, b), a(2−1, b)), i.e., δ(a(2, b), a(1, b)). In this case, the second sentence (n=2) is being compared to the first sentence (n=1). When calculating the variation degree for "South" (i.e., a(2, 5)), b=5 because "South" appears as the fifth word in the sentence. Thus, when "South" (a(2, 5)) is compared to "North" (a(1, 5)), δ(a(m, b), a(m−1, b)) in Expression 1 is "0" because a(2, 5) is not equal to a(1, 5). *See id*., ¶[0080] ("Further, in the mathematical expression (1), δ(a(m, b), a(m−1, b)) … represents '0' when a(m, b) is not equal to a(m−1, b).").

151. Second, the variable t(n) represents the time or order of the input:[7]

> ***[V]ariable t(n) represents the input*** or accumulation time ***assigned to the sentence specified by the variable n.*** For instance, t(1) represents the time when the sentence of [Today's weather in the Tohoku region is fair.] ([kyou no tohoku chihou no tenki wa hare desu]) is inputted or accumulated.

Id., ¶[0072].

152. Thus, "n" in Expression 1 is "2" because the second sentence (n=2) is being accumulated.

153. Third, as shown in the denominator of Expression 1, t(2)=2 and t(1)=1 because each represent the order of the input. For example, t(2) represents the second input (i.e., the second sentence (n=2)) and t(1) represents the first input (i.e., the first sentence (n=1)).

154. Accordingly, as shown below using Expression 1, the variation degree for "South" (a(2, 5)) is calculated to be "1."

---

[7] In the case where collation depends on "number of sentences" (*see* Ex. 1006 (*Matsumoto*), FIG. 3), t(n) indicates the sequential order in which the sentences are received, as show in figures 6-8. *See id*., ¶¶[0049], [0070]. If collation is based on time, t(n) represents the clock time at which the sentence was received (or spoken), as shown in figure 9. *See id*., ¶¶[0049], [0060], [0083].

$$v(\mathrm{n},\mathrm{b}) = \sum_{m=1}^{n} \frac{1-\delta(\mathrm{a(m,b)},\mathrm{a(m-1,b)})}{\mathrm{t(n)-t(m)+1}} = \frac{1-\delta(\mathrm{a(1,5)},\mathrm{a(0,5)})}{t(2)-t(1)+1} + \frac{1-\delta(\mathrm{a(2,5)},\mathrm{a(1,5)})}{t(2)-t(2)+1} =$$

$$\frac{1-1}{2-1+1} + \frac{1-0}{2-2+1} = 1$$

155. For the same reasons I discussed above, the variation degree for "Carolina" (a(2, 6)) is calculated to be "0" because "Carolina" (a(2, 6)) is the same as "Carolina" (a(1, 6)). More particularly, unlike the calculation for "South" described above, in this calculation, δ(a(m, b), a(m−1, b)) is "1" because a(2, 6) is equal to a(1, 6). *See* Ex. 1006 (*Matsumoto*), ¶[0080] ("[I]n the mathematical expression (1), δ(a(m, b), a(m−1, b)) represents '1' when a(m, b) is equal to a(m−1, b) … "). Therefore:

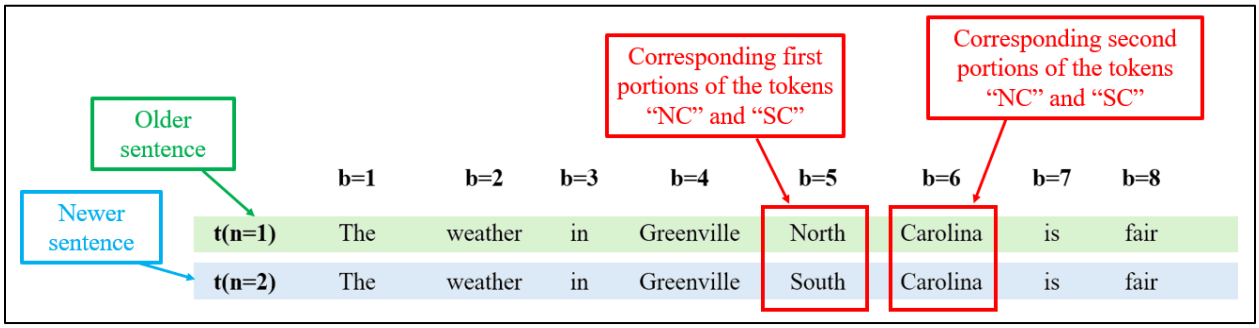
$$v(\mathrm{n},\mathrm{b}) = \sum_{m=1}^{n} \frac{1-\delta(\mathrm{a(m,b)},\mathrm{a(m-1,b)})}{\mathrm{t(n)-t(m)+1}} = \frac{1-\delta(\mathrm{a(1,6)},\mathrm{a(0,6)})}{t(2)-t(1)+1} + \frac{1-\delta(\mathrm{a(2,6)},\mathrm{a(1,6)})}{t(2)-t(2)+1} =$$

$$\frac{1-1}{2-1+1} + \frac{1-1}{2-2+1} = 0$$

156. Similar to "Carolina" as described above, the calculated variation degree for each of the remaining words in the second sentence (n=2) (i.e., a(2, 1), a (2, 2), a(2, 3), a(2, 4), a(2, 7) and a (2, 8)) are also calculated to be "0" because each word is determined to be the same as the corresponding word in sentence n=1. *See also id*., FIG. 6.

157. Further, in my opinion, a person of ordinary skill in the art would have understood that *Matsumoto*'s collation unit 14 compares ***corresponding***

portions of tokens because "South" and "North" both appear in the b=5 position and "Carolina" and "Carolina" both appear in the b=6 position.



158. In my opinion, a person of ordinary skill in the art would have understood that in the scenario discussed herein, "South" appearing in location a(2, 5) corresponds to "one first portion of at least one token" and "North" appearing in location a(1, 5) corresponds to "one corresponding first portion of at least one other token." Additionally, a person of ordinary skill in the art would have understood that "Carolina" appearing in location a(2, 6) corresponds to "one second portion of the at least one token" and "Carolina" appearing in location a(1, 6) corresponds to "one corresponding second portion of the at least one other token."

159. Indeed, the '291 patent describes comparing portions (i.e., "eleven" and "ten") of tokens (i.e., "11:45 a.m." and "10:45 a.m.") in a similar way. *See* Ex. 1001 (*'291Pat*), 28:27-31. The '291 patent states:

> In this example, "11:45 a.m.", the new time of the flight departure, contrasts with "10:45 a.m.", the original time of the flight departure. Specifically, ***"eleven" is the part of "11:45 a.m." that differs from and contrasts with the "ten" of "10:45 a.m." (i.e., the "forty_five" and the "a.m." do not differ or contrast).***

Id., 22:67-23:5.

> ***Identifying specific portions of one or more fields or tokens to be rendered to carry contrastive stress*** through a comparison of normalized orthography in some embodiments may provide the advantage that the portions that differ can be readily identified in terms of their spoken word forms as they will be rendered in the speech output.

Id., 28:62-67.

160. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach identifying a first portion (i.e., "North" and "South") of two tokens (i.e., "NC" and "SC") that are different and have a variation degree of "1" ("identifying at least one first portion of at least one token of the text input that differs from at least one corresponding first portion of at least one other token of the text input") and a second portion (i.e., "Carolina" and "Carolina") of two tokens (i.e., "NC" and "SC") that are the same and have a variation degree of "0" ("and at least one second portion of the at least one token that does not differ from at least one corresponding second portion of the at least one other token").

d. 1[c]: assigning contrastive stress to the identified at least one first portion of the at least one token, but not to the identified at least one second portion of the at least one token;

161. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

162. As I discussed in §XI.A.1.c(v) (application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report), in the example weather report discussed in this ground, "South" in location a(2, 5) corresponds to "one first portion of at least one token," and "North" in location a(1, 5) corresponds to "one corresponding first portion of at least one other token."

163. *Matsumoto* discloses "calculation unit 17" uses the "variation degree obtained from the collation unit 14" to "calculate[] the variation coefficients (the speed coefficient, the pitch coefficient and the volume coefficient)" between corresponding words or phrases of two or more sentences. Ex. 1006 (*Matsumoto*), ¶[0054].

164. The variation coefficients C1 (speed), C2 (pitch), and C3 (volume), are used to adjust the prosody of words/phrases in text when synthesized into speech. *See id.*, ¶¶[0103] ("the phoneme length generation unit 5 corrects the phoneme length with the ***speed coefficient*** (the variation coefficient) obtained from the coefficient calculation unit 10 (the calculation unit 17) …"), [0104] ("The pitch

generation unit 6 corrects (weights) the normal pitch frequency with the ***pitch coefficient*** obtained from the coefficient calculation unit 10 …"), [0105] ("the volume generation unit 7 multiples the normal volume by a ***volume coefficient*** obtained from the coefficient calculation unit 10 (the calculation unit 17), thereby correcting the volume …").

165. In my opinion, a person of ordinary skill in the art would have understood that the prosodic changes to "speed" recited in *Matsumoto* refers to changes in "duration" because it describes the speed of synthesized speech (i.e., how long it takes to speak a given word). *See id*., ¶[0010] (describing that the technology of a prior art reference discussed in *Matsumoto* "involves controlling the speech speed etc to make an output speech distinguishable in accordance with a result of the judgment."); *see also id*., ¶¶[0011]-[0014]. Further, *Matsumoto* explicitly correlates speed and phoneme length, such that speed is a scale factor for phoneme duration. *See id*., ¶[0103] ("For example, the phoneme length generation unit 5, when the phoneme length of a certain phoneme is 40 and the speed coefficient is 1.2, calculates a new phoneme length as 48.").

166. Further, in my opinion, a person of ordinary skill in the art would have understood that prosody is used to convey meaning in speech and aids in understanding a speaker.

167. For example, U.S. Pat. No. 7,313,523 ("*Bellegarda*") discloses a speech synthesis system that aims to "generat[e] variations in synthesized speech to produce speech that sounds more natural" Ex. 1014 (*Bellegarda*), 1:9-11. In particular, *Bellegarda* explains that "naturalness [of speech] affects intelligibility" and that "the more closely synthetic speech models natural speech, the more richly and redundantly the content and structure of the information will be represented in the acoustic signal. This in turn means that it will be easier for the listener to recover the intended meaning from the signal—i.e., the cognitive load associated with this task will be lower. Consequently, the task of understanding the speech will interfere less with other tasks the user is performing when using the computer system." *Id*., 1:40-53.[8]

168. Similarly, *Lewis* discloses a method for providing "more natural and realistic" speech using speech recognition markers (Ex. 1009 (*Lewis*), 2:53-58) because "widespread adoption of TTS technology has been prevented by the

[8] To be clear, I am referring to *Bellegarda* here as evidence of a person of ordinary skill in the art's knowledge and/or state of the art. While *Bellegarda* is part of Ground 1B, it is ***not*** part of Ground 1A.

perceived robotic quality of some voices and poor intelligibility of intonation-related cues" (*id*., 1:56-59).

169. Specifically, *Lewis* discloses a prosody control module 32 that uses prosody control rules 34 to assign frequency, duration, and amplitude to phonemic segments that are synthesized into speech. *See id*., 5:67-6:11. *Lewis* explains that "[p]rosody primarily involves pitch, duration, loudness, voice quality, tempo and rhythm." *Id*., 2:2-2:4.

170. Thus, in my opinion, a person of skill in the art would have understood that prosodic characteristics are used to convey meaning in speech because such characteristics are used to make speech sound more natural/realistic and therefore, more intelligible.

171. *Matsumoto* discloses that calculation unit 17 assigns the variation coefficients for speed/duration (C1), pitch (C2), and volume (C3) to the phoneme length generation unit 5, the pitch generation unit 6, and the volume generation unit 7, respectively. *See* Ex. 1006 (*Matsumoto*), ¶[0054] ("The calculation unit 17 assigns the speed coefficient to the phoneme length generation unit 5, assigns the pitch coefficient to the pitch generation unit 6 and assigns the volume coefficient to the volume generation unit 7.").

172. The variation coefficients are then used to adjust the speed/duration, pitch, and volume for each word or phrase of the new sentence. *See id*., ¶¶[0103]-[0105]; *see also id*., ¶¶[0042] ("The phoneme length generation unit 5, the pitch generation unit 6 and the volume generation unit 7 make, only when assigned a variation coefficient, the correction by use of the assigned variation coefficient."), [0088]-[0091].

173. *Matsumoto* further discloses a user can set a maximum and minimum value for the variation coefficients, which are upper and lower limits of the variation coefficients C1 (speed/duration), C2 (pitch), and C3 (volume):

> [T]he ***variation coefficient maximum value/minimum value setting unit 15 accepts an input, given by the user***, to the variation coefficient maximum value/minimum value setting window 28 through the input device (not shown) connected to the variation coefficient maximum value/minimum value setting unit 15 … ***An assumption is that the user inputs numerical values to the variation coefficient maximum value input field 29 and to the variation coefficient minimum value input field 30, and selects the setting button 31.***

Id., ¶¶[0064]-[0065].

174. *Matsumoto*'s figure 4 discloses an example in which the variation coefficients (C1, C2, and C3) have a user-set maximum value of 1.25 and minimum value of 0.85:

> [T]he variation coefficient is given a variation coefficient maximum value d(MAX) (which is 1.25 designated by the user in the present embodiment) and a variation coefficient minimum value d(MIN) (which is 0.85 designated by the user in the present embodiment), respectively.

Id., ¶[0091].

175. When calculation unit 17 calculates a variation coefficient that is more than the user-set maximum value, the maximum variation coefficient value is used:

> [I]f the calculated variation coefficient is larger than the variation coefficient maximum value d(MAX), the variation coefficient maximum value d(MAX) is adopted as a result of the calculation thereof.

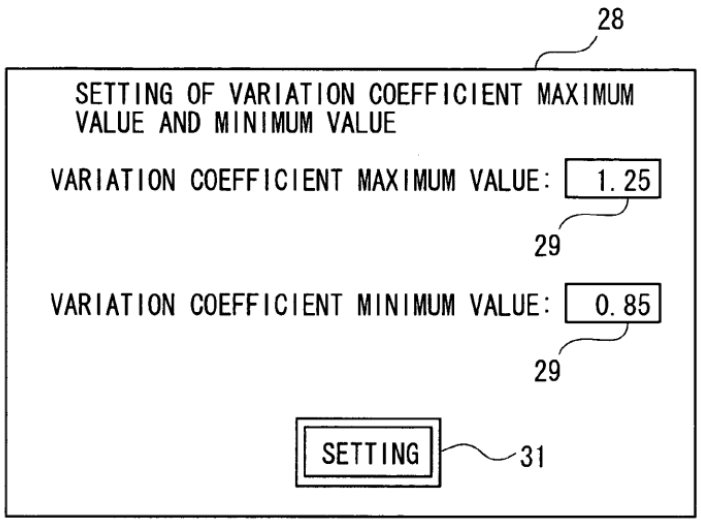
Id., ¶[0091].

176. Similarly, when calculation unit 17 calculates a variation coefficient that is less than the user-set minimum value, the minimum variation coefficient value is used:

> If the calculated variation coefficient is smaller than the variation coefficient minimum value d(MIN), the variation coefficient minimum value d(MIN) is adopted as a result of the calculation of the variation coefficient.

Id.

177. This is also shown in *Matsumoto*'s figure 4:



Ex. 1006 (*Matsumoto*), FIG. 4.

178. In my opinion, a person of ordinary skill in the art would have also understood that in *Matsumoto*'s system, a variation coefficient above 1.0 would emphasize a word/phrase, a variation coefficient below 1.0 would deemphasize a word/phrase, and a variation coefficient of 1.0 would neither emphasize nor deemphasize a word/phrase.

179. For example, a person of skill in the art would have understood that if the variation coefficient for the volume (C3) of synthesized speech is 1.25, the volume of the synthesized word would be increased because the normal volume is multiplied by the volume coefficient. *See* Ex. 1006 (*Matsumoto*), ¶[0105]. ("Subsequently, the volume generation unit 7 multiples the normal volume by a volume coefficient obtained from the coefficient calculation unit 10 (the

calculation unit 17), thereby correcting the volume (step S10).”). Thus, if the volume coefficient is greater than 1.0, the volume would increase when corrected.

180. Similarly, if the variation coefficient for volume was less than 1.0 (e.g., 0.85), then the volume would decrease, thus deemphasizing the word when synthesized. Finally, if the variation coefficient for volume was 1.0, then no prosodic changes would be implemented to the word when synthesized and the volume would neither increase nor decrease when corrected. *See id*., ¶[0112] (“In FIG. 13A, the speed (an original value) of the phoneme string is set to 1.0 in the case of executing none of the correction based on the speed coefficient.”).

181. To calculate the volume variation coefficient (C3) for “South” of the example weather report discussed herein, *Matsumoto* discloses using Expression 7, shown below. *See id*., ¶[0087] (“the volume coefficient is calculated by employing the following mathematical expression (7).”).[9]

[9] While I only discuss the calculation of the volume variation coefficient (C3), in my opinion, a person of ordinary skill in the art would have understood that the calculation of variation coefficients for speed/duration (C1) and pitch (C2) would be the same. *See* Ex. 1006 (*Matsumoto*), Expressions 5 and 6.

[Mathematical Expression 7]

$$C3(n, b) = \frac{v(n, b)ge(\mathrm{MIN})}{f \sum_{b=1} v(n, b)}$$

Ex. 1006 (*Matsumoto*), Expression 7.

182. First, as I previously described in §XI.A.1.c(v) (application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report), the variation degree for "South" (i.e., a(2, 5)) is calculated to be "1" because "South" is different from "North" (i.e., a (1, 5)). Thus, v(n, b) =1.

183. Second, as I discussed in §XI.A.1.c(v) (application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report), in the scenario discussed in this ground, each of the other words/phrases in the second sentence (n=2) has a calculated variation degree of "0" because they are identified to be the same as the corresponding word/phrase in the just anterior sentence (n=1). Therefore, $\sum_{b=1} v(n, b) = 1$.

184. Third, *Matsumoto* discloses that "g" is the "normal sentence length" while "f" is the normal phoneme length. *See id*., ¶[0090] ("the normal phoneme length f and the normal sentence length g"). In my opinion, a person of ordinary skill in the art would have understood that for "South," the normal phoneme length is 3 because "South" is made up of 3 phonemes (i.e., f=3). Similarly, the normal

sentence length is 32 because the second sentence (n=2) of the example weather report I have discussed here (i.e., "the weather in Greenville South Carolina is fair") comprises 32 phonemes (i.e., g=32). Thus, in Expression 7, g=32 and f=3.

185. Fourth, *Matsumoto* discloses that e(MIN) is "a preset coefficient minimum value e(MIN)." *Id*., ¶[0089]. While *Matsumoto* does not disclose what the present minimum coefficient would be set to, in my opinion, a person of ordinary skill in the art would have understood that e(MIN) operates as a scaling factor for the given coefficient and could be preset to a desired value. For example, the minimum preset value for the volume coefficient may be set to 0.75 such that the volume coefficient would not be lowered more than 25% from its original value. Thus, in my opinion, a person of ordinary skill in the art would have understood that e(MIN) for the volume variation coefficient (C3) may be set to 0.75 to reduce the volume coefficient by a factor of 0.75.

186. Using each of the values I described above, C3 for "South" (a(2, 5)) is calculated to be 8 as shown below.

$$C3(n,b) = \frac{v(n,b)ge(MIN)}{f\sum_{b=1} v(n,b)} = \frac{(1)(32)(0.75)}{(3)(1)} = 8$$

187. Further, as shown in figure 4, the preset user maximum variation coefficient value is 1.25 while the minimum variation coefficient value is 0.85. *See* Ex. 1006 (*Matsumoto*), FIG. 4. Because C3 is calculated to be 8, and therefore,

above the user-set maximum coefficient value of 1.25, the maximum variation coefficient of 1.25 is adopted. *See id*., ¶[0091] ("[I]f the calculated variation coefficient is larger than the variation coefficient maximum value d(MAX), the variation coefficient maximum value d(MAX) is adopted as a result of the calculation thereof.").

188. Accordingly, the variation coefficient for volume (C3) for "South" is set to 1.25. As such, the volume for "South" would be increased by 25% when synthesized into speech.

189. Additionally, as I discussed above, when calculation unit 17 calculates a variation coefficient that is less than the user-set minimum value, the minimum variation coefficient value is used. *See id*., ¶[0091]. Thus, when the variation degree is calculated to be "0" (i.e., words are determined to be the same), the variation coefficient would also be calculated to be "0."

190. For example, if the variation degree (i.e., v(n, b)) for a given word is "0," the entire variation coefficient for volume (C3) would also be calculated to be "0." To illustrate, the variation coefficient for "Carolina" (a(2, 6)), is calculated to be "0" as shown below.

$$C3(n,b) = \frac{v(n,b)ge(MIN)}{f\sum_{b=1} v(n,b)} = \frac{(0)(32)(0.75)}{(8)(1)} = 0$$

191. Further, because C3 is calculated to be 0, and therefore, below the user-set maximum coefficient value of 0.85 (*see id*., FIG. 4), the minimum variation coefficient of 0.85 is adopted. *See id*., ¶[0091].

192. It is my opinion that a person of ordinary skill in the art would have understood that a user would not preset the minimum variation coefficient value shown in figure 4 to "0" because this would result in each of the coefficients for speed/duration (C1), pitch (C2), and volume (C3) to be "0" when words are calculated to be the same. Further, if each of the variation coefficients for speed/duration, pitch, and volume were "0," no synthesized speech would be produced as the corrected phonemes would have zero duration and no volume. Therefore, it is my opinion that a person of ordinary skill in the art would set a minimum variation coefficient value to a number above "0" (e.g., 0.75, 0.85, 0.95, 1.0, etc.).

193. Thus, in the example weather report discussed herein, in which only the first two sentences are collated (i.e., n=2 and n=1), the volume variation coefficient for "South" (a(2, 5)) is 1.25 while the volume variation coefficient for "Carolina" (a(2, 6)) is 0.85. Thus, when corrected, the volume for "South" is increased, while the volume for "Carolina" is decreased. Accordingly, only the

word determined to be different between two text strings (i.e., "South") is emphasized while words that are the same (e.g., "Carolina") are deemphasized.

194. Nevertheless, in my opinion, a person of ordinary skill in the art would have understood that deemphasizing words or phrases that are the same (e.g., by decreasing the pitch, amplitude and/or speed/duration), would still constitute "***not*** [assigning contrastive stress] to the identified at least one second portion of the at least one token," as recited in claim limitation 1[c] and understood in the context of the '291 patent.

195. The '291 patent describes "contrastive stress" as ***increasing*** the prosody for the words/phrases meant to contrast with each other. *See* Ex. 1001 (*'291Pat*), 9:22-27 ("[S]ynthesis system may apply the ***contrastive stress to the identified word(s) and/or syllable(s) through increased pitch, amplitude and/or duration,*** or in any other suitable manner."), 12:44-48 ("***[C]ontrastive stress may be a stress or emphasis of a greater degree*** than would normally be applied to the same word when it is not being distinguished or contrasted with another word of like type, function and/or subject matter.").

196. The '291 patent does not disclose or otherwise suggest that decreasing the pitch, amplitude, and/or duration of words/phrases (e.g., to deemphasize a word/phrase) constitutes applying contrastive stress.

197. Additionally, during prosecution of U.S. Patent 8,447,610 ("the '610 patent"), Applicant stated that "'[c]ontrastive stress' is a term of art, whose meaning is explained in detail in the specification, e.g., at pages 11-12."[10] Ex. 1013 (*'610FH*), 83.

198. The portion of the specification cited by the Applicant describes contrastive stress as "a particular emphasis pattern [] applied in speech to words or syllables that are ***meant to contrast with each other***." Ex. 1013 (*'610FH*), 147; Ex. 1018 (*'610Pat*), 8:12-15; Ex. 1001 (*'291Pat*), 8:18-20. That portion of the specification further states that "[h]uman speakers naturally apply contrastive stress to ***emphasize*** a word or syllable contrary to its normal accentuation, in order to contrast it with an alternative word or syllable or to focus attention on it." Ex. 1013 (*'610FH*), 147; Ex. 1018 (*'610Pat*), 8:15-18; Ex. 1001 (*'291Pat*), 8:20-23.

199. Moreover, Applicant explicitly rejected the Examiner's suggestion that contrastive stress could be read on any form of stress that differs from one word and any other word:

[10] I understand that the '610 patent is in the same family as the '291 patent (*see* Ex. 1022). I have been informed by counsel that I can consider the file history of the '610 patent.

> The Examiner questioned whether the term "contrastive stress" could be read on any form of stress that differs between one word and any other word. The Assignee's representatives pointed out that such an interpretation would not be consistent with how the term "contrastive stress" is generally understood in the art. ***Contrastive stress is a particular type of stress, and not every type of stress is contrastive stress***.

Ex. 1013 (*'610FH*), 83.

200. Accordingly, in my opinion, both the specification and prosecution history of the '291 patent are consistent with a person of ordinary skill in the art's understanding that deemphasizing words that are the same is not an application of contrastive stress.

201. Therefore, in my opinion, a person of ordinary skill in the art would have understood that when *Matsumoto* deemphasizes words or phrases that are the same, *Matsumoto* does not assign "contrastive stress" because *Matsumoto* is not applying a stress to words or phrases ***meant to contrast*** with each other.

202. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach assigning stress by increasing the speed/duration, pitch, and volume using variation coefficients C1, C2, and C3 to the identified first portion of the at least one token (e.g., "South" of "SC") of a weather report ("assigning contrastive stress to the identified at least one first portion of the at least one token") but not increasing the

speed/duration, pitch, and volume for the second portion of the at least one token (e.g., "Carolina" of "SC") ("but not to the identified at least one second portion of the at least one token").

e. 1[d]: generating, using at least one computer system, an audio speech output corresponding to at least a portion of the text input, the audio speech output comprising at least one portion carrying contrastive stress corresponding to the at least one first portion of the at least one token of the text input, to contrast with at least one other portion of the audio speech output corresponding to the at least one corresponding first portion of the at least one other token of the text input; and

203. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

204. *Matsumoto* discloses a speech synthesizer that generates a synthesized speech output for a given text input by adjusting the speed/duration, pitch, and volume of words or phrases of a new sentence based on assignment of contrastive stress using coefficients C1 (speed/duration), C2 (pitch), and C3 (volume):

> [T]he phoneme length generation unit 5 ***corrects the phoneme length with the speed coefficient*** (the variation coefficient) obtained from the coefficient calculation unit 10 (the calculation unit 17) …

Ex. 1006 (*Matsumoto*), ¶[0103].

> The pitch generation unit 6 ***corrects (weights) the normal pitch frequency with the pitch coefficient*** obtained from the coefficient calculation unit 10 …

Id., ¶[0104].

> [T]he volume generation unit 7 ***multiples the normal volume by a volume coefficient*** obtained from the coefficient calculation unit 10 (the calculation unit 17), thereby correcting the volume …

Id., ¶[0105].

205. Specifically, *Matsumoto* discloses the phoneme string with changes to speed/duration, pitch and volume is generated by the "waveform generation unit." *Id*., ¶[0107]; *see also id*., ¶[0043] ("The waveform generation unit 8 generates a synthetic speech from the phoneme length, the phoneme string, the pitch pattern and the volume information by a predetermined method, and outputs the synthetic speech.").

206. In my opinion, each of *Walker Matsumoto*, and *Malsheen* disclose that their inventions are performed using a computer-based system.

207. For example, *Walker* discloses that its "TTS engines … synthesize the written words of a text document into ***computer generated*** speech understandable by humans." Ex. 1005 (*Walker*), ¶[0003]; *see also id*., ¶¶[0019] (disclosing that *Walker* is "critical for wider acceptance of ***voice-computer interfaces***"), [0024]

(Walker's TTS engine "converts or synthesizes … text from voice application 16 into ***computer generated audio speech***.").

208. Similarly, *Matsumoto* discloses that "[t]he present invention can be actualized as a program that ***makes a computer function as the speech synthesizer***." Ex. 1006 (*Matsumoto*), ¶[0018]. *Matsumoto* further discloses:

> ***The speech synthesizer 1 can be actualized by use of a hard disc (storage device) storing a program*** for executing processes in the embodiment executed, ***a central processing unit (CPU) that executes this program*** and a computer (information processing device) having a memory employed for temporarily storing information, and the configuration described above is a function actualized in such a way that the CPU loads the program stored in the hard disc into the memory and executes this program.")

Id., ¶[0036].

> FIG. 10 is a flowchart showing an operating example (processing example) of the speech synthesizer 1. When a power source of the speech synthesizer 1 is switched ON, the ***central processing unit (CPU) provided in the speech synthesizer 1 reads a program for generating the synthetic speech from the hard disc (storage device), then loads the program into the memory and executes the program***. Through this operation, a process shown in FIG. 10 comes to a start-enabled status. The start of the process shown in FIG. 10 is triggered by inputting the text data for generating the synthetic speech to the input unit 3.

Id., ¶[0093].

209. *Malsheen* discloses that "the present invention is a ***computer system*** for converting a text message into an audio representation … ." Ex. 1007 (*Malsheen*), 2:62-65; *see also id*., 3:51-4:5, 4:17-35.

210. Thus, in my opinion, a person of ordinary skill in the art would have understood that the *Walker-Matsumoto-Malsheen* combination's functionality, including generating synthesized speech with appropriate contrastive stress, would also be implemented using a computer system.

211. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach using a computer system to generate audio speech output of a text weather report ("generating, using at least one computer system, an audio speech output corresponding to at least a portion of the text input") with changes to speed/duration, pitch and volume using the assigned coefficients C1, C2, and C3 for a differing portion of a token (e.g., "South" of "SC") of a new sentence (n=2) ("the audio speech output comprising at least one portion carrying contrastive stress corresponding to the at least one first portion of the at least one token of the text input") but not adjusting the speed/duration, pitch, and volume for a same portion of a token (e.g., "North" of "NC") in a prior sentence (n=1) ("to contrast with at least one other portion of the audio speech output corresponding to the at

least one corresponding first portion of the at least one other token of the text input").

f. 1[e]: providing the audio speech output for the speech-enabled application.

212. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

213. *Walker* discloses that "TTS engine 18 … converts the cleaned text into speech and provides the speech which is labeled 'Synthesized Audio' in FIG. 1 to voice application 16." Ex. 1005 (*Walker*), ¶[0034].

214. As I discussed in §XI.A.1.c(ii) (motivation to combine *Walker* and *Matsumoto*), it is my opinion that a person of ordinary skill in the art would have found it obvious to incorporate *Matsumoto*'s TTS system into *Walker*'s TTS engine farm as the TTS engine for synthesizing text weather reports into speech.

215. As I discussed in §XI.A.1.a (limitation 1[pre]), it is my opinion that a person of ordinary skill in the art would have understood that voice application 16 is a speech-enabled application.

216. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach a TTS engine providing synthesized speech to a voice application ("providing the audio speech output for the speech-enabled application").

2. Independent Claim 8

a. 8[pre]: Apparatus for providing speech output for a speech-enabled application, the apparatus comprising:

217. I understand that "[a]pparatus for providing speech output for a speech-enabled application, the apparatus comprising" is the preamble to claim 8 of the '291 patent. I have been asked to assume that the preamble is a claim limitation. Under that assumption, in my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches the preamble.

b. 8[a]: a memory storing a plurality of processor-executable instructions; and 8[b]: at least one processor, operatively coupled to the memory, that executes the instructions to:

218. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

219. As I discussed in §XI.A.1.e (limitation 1[d]) each of *Walker*, *Matsumoto*, and *Malsheen* disclose that their inventions are performed using a computer-based system. As also discussed in §§XI.A.1.c(ii) (motivation to combine *Walker* and *Matsumoto*) and XI.A.1.c(iv) (motivation to combine *Walker-Matsumoto* with *Malsheen*), it is my opinion that a person of ordinary skill in the art would have found it obvious to combine the teachings of *Walker*, *Matsumoto*, and *Malsheen*.

220. In my opinion, a person of ordinary skill in the art would have found it obvious for the *Walker-Matsumoto-Malsheen* combination to include a memory for storing a plurality of processor-executable instructions, and to also include a processor, operatively couple to the memory, to execute the instructions because at least each of *Matsumoto* and *Malsheen* explicitly disclose such an arrangement:

> The present invention can be actualized as a synthetic speech generation method having the same features as those of the speech synthesizer described above. Further, the present invention can be actualized as a program that makes a computer function as the speech synthesizer described above and as a ***storage medium storing this program.***

Ex. 1006 (*Matsumoto*), ¶[0018].

> The speech synthesizer 1 can be actualized by use of a hard disc (storage device) storing a program for executing processes in the embodiment executed … and the configuration described above is a function actualized in such a way that ***the CPU loads the program stored in the hard disc into the memory and executes this program***.

Id., ¶[0036].

> … the ***central processing unit (CPU)*** provided in the speech synthesizer 1 ***reads a program for generating the synthetic speech from the hard disc (storage device)***, then loads the program into the memory and executes the program.

Id., ¶[0093]; *see also* FIG. 10.

> The computer system 100 includes ***a central processing unit*** 102 … The ***computer memory 106 stores a number of different programs, sometimes herein called procedures, and data structures***. These programs and data structures, which are discussed in more detail below, include … a text reader module 128.

Ex. 1007 (*Malsheen*), 3:59-4:5, FIG. 2.

221. Further, in my opinion, a person of ordinary skill in the art would have also found it obvious to combine the teachings of *Walker*, *Matsumoto*, and *Malsheen* because it amounted to use of a known technique (*Matsumoto* and *Malsheen* each teach a computer having a processor operatively coupled to a non-transitory memory storing instructions that are executed by the processor) to improve a similar device (the *Walker-Matsumoto-Malsheen* combination is implemented using a computer) in the same way (incorporating a processor coupled to non-transitory memory storing instructions that are executed by the processor in the *Walker-Matsumoto-Malsheen* combination's computer).

222. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach a computer that includes a hard disk or other memory that stores processor-executable instructions ("a memory storing a plurality of processor-executable instructions") to be carried out by the computer's processor ("at least one processor

… that executes the instructions"), wherein the processor reads the instructions from the memory ("operatively coupled to the memory").

c. 8[c]-[g]

223. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches Elements 8[c]-8[g] for the same reasons discussed in §§XI.A.1.b (limitation 1[a]) - XI.A.1.f (limitation 1[e]), according to the table below.

Element	Section
8[c]	XI.A.1.b (limitation 1[a])
8[d]	XI.A.1.c(i) through XI.A.1.c(v) (limitation 1[b])
8[e]	XI.A.1.d (limitation 1[c])
8[f]	XI.A.1.e (limitation 1[d])
8[g]	XI.A.1.f (limitation 1[e])

3. Independent Claim 14

a. 14[Pre]: At least one non-transitory computer-readable storage medium encoded with a plurality of computer-executable instructions that, when executed, perform a method for providing speech output for a speech-enabled application, the method comprising:

224. I understand that "[a]t least one non-transitory computer-readable storage medium encoded with a plurality of computer-executable instructions that, when executed, perform a method for providing speech output for a speech-

enabled application, the method comprising" is the preamble to claim 14 of the '291 patent. I have been asked to assume that the preamble is a claim limitation. Under that assumption, in my opinion, *Walker* alone or in combination with *Matsumoto* and *Malsheen* teaches the preamble.

225. For the reasons I discussed in §XI.A.1.a (limitation 1[pre]), *Walker* teaches "a method for providing speech output for a speech-enabled application."

226. *Walker* discloses that "[t]ext-to speech (TTS) engines are ***computing devices*** which convert written text into audible ***computer generated speech***." Ex. 1005 (*Walker*), ¶[0003]; *see also id*., ¶¶[0019] (disclosing that *Walker* is "critical for wider acceptance of ***voice-computer interfaces***"), [0024] (*Walker*'s TTS engine "converts or synthesizes … text from ***voice application 16*** into ***computer generated*** audio speech.").

227. Accordingly, in my opinion, a person of ordinary skill in the art would have understood that *Walker*'s method is performed using software (i.e., computer readable instructions) running on a computer.

228. This understanding is confirmed, for example, by the fact that the provisional application from which *Walker* claims priority refers to the context detector and text cleaner as "software components." Ex. 1015 (*'000Prov.*), 5.

229. A person of ordinary skill in the art would have further understood that *Walker*'s computer comprises at least one processor for executing the software to implement *Walker*'s method, as evidenced, for example, by *Bellegarda*. *See* Ex. 1014 (*Bellegarda*). *Bellegarda* is directed to a speech synthesis system and its disclosure is analogous to that of *Walker* and *Matsumoto*. *Bellegarda* discloses:

> The methods to be performed by a computer constitute computer programs made up of computer-executable instructions. Describing the methods by reference to a flowchart enables one skilled in the art to develop such programs including such instructions to carry out the methods on suitably configured computers (the ***processor of the computer executing the instructions from computer-accessible media***) … Furthermore, it is common in the art to speak of software, in one form or another (e.g., program, procedure, process, application …), as taking an action or causing a result. Such expressions are merely a shorthand way of saying that ***execution of the software by a computer causes the processor of the computer to perform an action or a produce a result***."

Ex. 1014 (*Bellegarda*), 10:8-31.[11]

230. Additionally, in my opinion, a person of ordinary skill in the art would have understood that such software components/instructions would be stored in a

[11] *Bellegarda* is not part of Ground 1A. *See* footnote 8.

non-transitory computer-readable storage medium to be available for execution (e.g., so as to not be lost when the computer is powered off).

231. To the extent Patent Owner argues and the Board finds that *Walker* alone does not teach a non-transitory computer-readable storage medium for storing the software components/instructions executed by *Walker*'s computing device, it is my opinion that this feature would have been obvious based on *Walker* in combination with *Matsumoto* and *Malsheen*.

232. For example, *Matsumoto* discloses that its TTS system "can be actualized as a program that makes a computer function as the speech synthesizer … and as a ***storage medium storing this program***." Ex. 1006 (*Matsumoto*), ¶[0018]; *see also id*., ¶¶[0036] ("The speech synthesizer 1 can be actualized by use of a hard disc (storage device) storing a program for executing processes in the embodiment executed … and the configuration described above is a function actualized in such a way that ***the CPU loads the program stored in the hard disc into the memory and executes this program***."), [0093] ("the ***central processing unit (CPU)*** provided in the speech synthesizer 1 ***reads a program for generating the synthetic speech from the hard disc (storage device)***, then loads the program into the memory and executes the program."), FIG. 10. Thus, in my opinion, a person of ordinary skill in the art would have found it obvious to implement the

functionality of the *Walker-Matsumoto-Malsheen* combination as computer-executable instructions stored on a non-transitory computer-readable medium because *Matsumoto* expressly teaches to do so.

233. Further, it is my opinion that a person of ordinary skill in the art would have also found it obvious to do so because it amounted to use of a known technique (*Matsumoto* teaches a non-transitory memory for storing program instructions executed by a computer) to improve a similar device (*Matsumoto* teaches methods performed by a computer) in the same way (incorporating a non-transitory memory in the *Walker-Matsumoto-Malsheen* combination to store the program instruction executed by the computer).

234. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach a system that includes a hard disk or other non-transitory memory ("at least one non-transitory computer readable medium") that stores computer-executable code to be carried out by a computer's processor ("storing instructions that, when executed by at least one processor") to implement the conversion of text from a voice application into speech ("that, when executed, perform a method for use with a speech-enabled application").

b. 14[a]-[e]

235. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches Elements 14[a]-14[e] for the same reasons I discussed in §§XI.A.1.b (limitation 1[a]) - XI.A.1.f (limitation 1[e]), according to the table below.

Element	Section
14[a]	XI.A.1.b (limitation 1[a])
14[b]	XI.A.1.c(i) through XI.A.1.c(v) (limitation 1[b])
14[c]	XI.A.1.d (limitation 1[c])
14[d]	XI.A.1.e (limitation 1[d])
14[e]	XI.A.1.f (limitation 1[e])

4. Dependent Claims 2-4, 6-7, 9-11, 13, 15-17 and 19-20

a. Claim 2

(i) 2[a]: The method of claim 1, wherein the identifying comprises: identifying a plurality of tokens of the text input of a same text normalization type for which a contrastive stress pattern is to be applied;

236. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

237. The ’291 patent describes “the same text normalization type” as encompassing the same word type, including state names:

> ***Examples of text normalization types*** of text input fields to which synthesis system 200 may apply contrastive stress patterns include, but are not limited to, alphanumeric sequence, address, Boolean value (true or false), currency, date, digit sequence, fractional number, proper name, number, ordinal number, telephone number, flight number, ***state name***, street name, street number, time and zipcode types.

Ex. 1001 (*'291Pat*)., 24:45-51.

238. As I discussed in §XI.A.1.c(iv) (motivation to combine *Walker*, *Matsumoto*, and *Malsheen*), *Malsheen* discloses a "text classifier procedure" that classifies whether a word is a state name. Classifying an abbreviation based on type (e.g., state name) is consistent with how the '291 patent describes identifying the same text normalization type of the input text:

> A tokenizer of synthesis system 200 may also be programmed, in accordance with some embodiments of the present invention, to ***analyze the tokens that it parses to infer their text normalization type***. For example, a tokenizer of synthesis system 200 may be programmed to determine that ***"9:42 a.m.", "10:30 a.m.", "11:30 a.m." and "12:30 p.m." in text input 350 are of the "time" text normalization type*** based on their syntax (e.g., one or two numerals, followed by a colon, followed by two numerals, followed by "a.m." or "p.m.").

Id., 32:7-16.

239. For the reasons I discussed in §§XI.A.1.c(i) ("token" as used in the '291 claims and specification) and XI.A.1.c(iv) (motivation to combine *Walker*,

Matsumoto, and *Malsheen*), a person of ordinary skill in the art would have understood the abbreviations "NC" and "SC" in the example weather report discussed in this ground to comprise "tokens."

240. As I also discussed in §XI.A.1.c(iv) (motivation to combine *Walker*, *Matsumoto*, and *Malsheen*), in the *Walker-Matsumoto-Malsheen* combination, the context cleaner would recognize each of "NC" and "SC" to correspond to an abbreviation of a state name and would expand the abbreviations to the full name of the state (e.g., "North Carlina" and "South Carolina").

241. Additionally, as I discussed in §XI.A.1.d (limitation 1[c]), the *Walker-Matsumoto-Malsheen* combination assigns a contrastive stress pattern to "South," which is a portion of the token "NC."

242. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach identifying that each of the abbreviations "NC" and "SC" ("identifying a plurality of tokens") represent state names ("of a same text normalization type") and assigns contrastive stress pattern to "South" which is portion of the expanded token "SC" ("for which a contrastive stress pattern is to be applied").

243. Further, as described in §§XI.A.1.c(v) (application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report) and

XI.A.1.d (limitation 1[c]), *Matsumoto*'s collation unit operates on the expanded state names "North Carolina" and "South Carolina" (i.e., tokens that have undergone text normalization) and identifies the word "South" is different from "North" and therefore, "South" is assigned contrastive stress.

244. Further, the '291 patent contemplates that applying a "contrastive stress pattern" to a plurality of tokens includes assigning contrastive stress to one of two tokens, and in some embodiments, a portion of one of two tokens. *See* Ex. 1001 (*'291Pat*), 25:31-34 ("In some other embodiments, synthesis system 200 may be programmed to apply a ***contrastive stress pattern in which only the second of two contrasting fields (i.e., '1145A') is rendered with stress.***"); 23:3-5 ("Specifically, 'eleven' is the part of '11:45 a.m.' that differs from and contrasts with the 'ten' of '10:45 a.m.' (i.e., the 'forty_five' and the 'a.m.' do not differ or contrast)").

245. Thus, in my opinion, a person of ordinary skill in the art would have understood *Matsumoto*'s collation unit 14 identifies one token (i.e., "SC") of the plurality of tokens (i.e., "NC" or "SC") of the same text normalization type (i.e., state name) to receive contrastive stress (i.e., "South" receives prosodic adjustment) when synthesized into speech.

246. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach identifying a same text normalization type (i.e., state names) for a plurality of tokens (i.e., "NC" and "SC") ("wherein the identifying comprises: identifying a plurality of tokens of the text input of a same text normalization type for which a contrastive stress pattern is to be applied") and identifying one token (i.e., "SC") to receive contrastive stress from the plurality of tokens (i.e., "NC" or "SC") of the same text normalization type (i.e., state names) ("and identifying the at least one token to be rendered with contrastive stress from among the plurality of tokens of the same text normalization type").

(ii) 2[b]: identifying the at least one token to be rendered with contrastive stress from among the plurality of tokens of the same text normalization type.

247. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

248. The '291 patent contemplates that applying a "contrastive stress pattern" to a plurality of tokens includes assigning contrastive stress to one of two tokens, and in some embodiments, a portion of one of two tokens: *See* Ex. 1001 (*'291Pat*), 25:31-34 ("In some other embodiments, synthesis system 200 may be programmed to apply a ***contrastive stress pattern in which only the second of two***

contrasting fields (i.e., '1145A') is rendered with stress."); 23:3-5 ("Specifically, ***'eleven' is the part of '11:45 a.m.' that differs from and contrasts with the 'ten' of '10:45 a.m.'*** (i.e., the 'forty_five' and the 'a.m.' do not differ or contrast)").

249. Thus, in my opinion, a person of ordinary skill in the art would have understood *Matsumoto*'s collation unit 14 identifies one token (i.e., "SC") of the plurality of tokens (i.e., "NC" or "SC") of the same text normalization type (i.e., state name) to receive contrastive stress (i.e., "South" receives prosodic adjustment) when synthesized into speech.

250. For the reasons I discussed in §§XI.A.1.c(v) (application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report)and XI.A.1.d (limitation 1[c]), the *Walker-Matsumoto-Malsheen* combination teaches identifying "South" as the portion of the abbreviation "SC" after it has been expanded to "South Carolina" to which contrastive stress is assigned ("identifying the at least one token to be rendered with contrastive stress") from among "NC" and "SC" both of which represent abbreviations of state names ("from among the plurality of tokens of the same text normalization type").

251. This is consistent with how the '291 patent describes assigning contrastive stress to a portion of a token after it has undergone normalized orthography:

> In some embodiments, ***synthesis system 200 may identify the specific sub-portion(s) of one or more fields or tokens that should carry contrastive stress by comparing the normalized orthography of the field(s) or token(s)*** identified to be rendered with contrastive stress to the normalized orthography of the other contrasting field(s) or token(s), and determining which specific portion or portions differ.

Ex. 1001 (*'291Pat*), 27:66-28:5.

> [I]n the example of FIG. 3A, ***synthesis system 200 may compare the normalized orthography "eleven_forty_five_a_m" to the normalized orthography "ten_forty_five_a_m_" to determine that "eleven" is the portion that differs***. As a result, synthesis system 200 may assign the word "eleven" to carry contrastive stress.

Id., 28:27-32.

> ***Identifying specific portions of one or more fields or tokens to be rendered to carry contrastive stress through a comparison of normalized orthography*** in some embodiments may provide the advantage that the portions that differ can be readily identified in terms of their spoken word forms as they will be rendered in the speech output.

Id., 28:62-67.

252. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach identifying "South" as the portion of the abbreviation "SC" after it has been expanded to "South Carolina" to which contrastive stress is assigned ("identifying

the at least one token to be rendered with contrastive stress") from among "NC" and "SC" both of which represent abbreviations of state names ("from among the plurality of tokens of the same text normalization type").

b. Claim 3: The method of claim 2, wherein the same text normalization type is a text normalization type selected from the group consisting of: an alphanumeric sequence type, an address type, a Boolean value type, a currency type, a date type, a digit sequence type, a fractional number type, a proper name type, a number type, an ordinal number type, a telephone number type, a flight number type, a state name type, a street name type, a street number type, a time type and a zipcode type.

253. For the same reasons discussed in §XI.A.4.a (Claim 2) , *Walker-Matsumoto-Malsheen* combination teaches this claim, particularly that "the same text normalization type is a text normalization type selected from the group consisting of: … a state name type."

c. Claim 4: The method of claim 2, wherein the plurality of tokens are identified based at least in part on at least one indication in the text input that the contrastive stress pattern is desired in association with the plurality of tokens.

254. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this claim.

255. As discussed in §§XI.A.1.c (limitation 1[b]) - XI.A.1.d (limitation 1[c]), *Matsumoto* discloses comparing words in corresponding "b" positions of a

current sentence and a prior sentence to identify words that are different and then adjusting the pitch, speed, and volume of the words that are different (i.e., assigns "contrastive stress" to the words that are different).

256. Comparing corresponding words (i.e., in the same "b" position) is consistent with how the '291 patent describes indications in the text input of contrastive stress. *See* Ex. 1001 (*'291Pat*), 26:52-55 ("In some embodiments, fields and/or tokens to be rendered with contrastive stress may also or alternatively be identified based on part-of-speech sequences that establish a repeated pattern with one element different.").

257. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach identifying the abbreviations "NC" and "SC" representing state names ("wherein the plurality of tokens are identified"), expanding the abbreviations to the full state name and assigning stress to the word "South" in the new sentence based on the fact that it is different from the worth "North" in the same "b" position in the previous sentence ("based at least in part on at least one indication in the text input that the contrastive stress pattern is desired in association with the plurality of tokens").

d. Claim 6

(i) 6[a]: The method of claim 2, wherein identifying the plurality of tokens comprises: tokenizing the text input;

258. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this feature.

259. In my opinion, a person of ordinary skill in the art would have understood the term "tokenizing" or "tokenization" to refer to breaking up input text into its component parts (e.g., words/tokens).

260. For example, *Speech Synthesis*, in *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Daniel Jurafsky & James H. Martin Oct. 13, 2007) ("*Jurafsky*") is a well-known publication that discusses speech and language processing and provides insight into how speech synthesizers operate. *See* Ex. 1016 (*Jurafsky*). *Jurafsky* describes "tokenization" as "segment[ing] [a] paragraph into separate utterances for synthesis …" (*id.*, at 3) and as "separate[ing] out and identify[ing] potential non-standard words" (*id.*, at 6). *Jurafsky* also describes tokenization as breaking up text based on whitespace:

> In the ***tokenization step, we can tokenize the input by whitespace*** …
> Tokenization algorithms also need to split words which are combinations of

two tokens, like 2-car or RVing. Words can be split by simple heuristics, such as splitting at dashes, or at changes from lower-case to upper-case.

Id., at 6.

261. Furthermore, U.S. Pat. No. 6,865,533 ("*Addison*") discloses a text to speech system in which text is examined and marked-up according to phonetic markup systems. Ex. 1011 (*Addison*), 6:52-7:9. Indeed, text is first tokenized before it is processed and synthesized:

> The ***process of tokenization constitutes producing a word list for input text*** in a file or record being transformed into speech in accordance with the present invention. For example, in the question: 'Mr. Smith, are you going to New York on June 5?', the output of the first part of the tokenizing operation appears as:
>
> Mr., Smith, [comma], are, you, going, to, New, York, on, June, 5, [?].

Id., 8:15-23.

262. Similarly, U.S. Pat. No. 7,725,321 ("*Bennett*") discloses an "intelligent query system for processing voiced-based queries" (Ex. 1012 (*Bennett*), Abstract) that includes a speech recognition system and natural language processor that tokenizes input text before it is synthesized into speech. *See id*., 19:47-20:42. *Bennett* further describes tokenization as breaking text into meaningful units that are larger than characters, but smaller than phrases/sentences:

> ***Tokenization is implemented by a text analyzer which treats the text as a series of tokens or useful meaningful units that are larger than individual characters, but smaller than phrases and sentences. These include words, separable parts of words, and punctuation.*** Each token is associated with an offset and a length. The first phase of tokenization is the process of segmentation which extracts the individual tokens from the input text and keeps track of the offset where each token originated in the input text.

Id., 20:23-31.

263. The ’291 patent uses the term “tokenizing” consistent with this understanding:

> Synthesis system 200 may be programmed, e.g., through a tokenizer implemented as part of front-end 250, to parse text input 350 into a sequence of ***individual tokens on the order of single words***.

Ex. 1001 (*’291Pat*), 31:56-59.

> At act 410, the text input may be ***tokenized, i.e., parsed into individual tokens on the order of single words***.”

Id., 36:23-25; *see also id*., 32:2-6 (“It should be appreciated, however, that synthesis system 200 may be programmed to tokenize text input 350 using any suitable tokenization technique in accordance with any suitable tokenization rules and/or criteria, as aspects of the present invention are not limited in this respect.”).

264. *Walker* discloses that “context detector 26 processes the raw text to determine the context of the text by locating in the text key words and phrases

associated with each context detection rule set stored in context detection rules database 21" Ex. 1005 (*Walker*), ¶[0030]; *see also id*., ¶[0031].

265. In my opinion, a person of ordinary skill in the art would have understood that for *Walker*'s context detector to compare words in the raw text input with words in a context detection rules databases, the raw text input would have been broken up into its constituent words for analysis. In other words, it is my opinion that a person of ordinary skill in the art reading *Walker* would have understood *Walker* to teach breaking the raw text input into its individual words, and then comparing those individual words to words in the context detection rules database.

266. To the extent Patent Owner argues and the Board finds that *Walker* alone does not teach tokenizing the text input, doing so would have been obvious based on *Malsheen*'s teachings. For the reasons I discussed in §XI.A.1.c(iv) (motivation to combine *Walker*, *Matsumoto*, and *Malsheen*), it would have been obvious to a person of ordinary skill in the art to incorporate *Malsheen*'s text classifier procedure that determines and classifies the syntax aspects of each word including making determinations as to whether a word is a member of a particular class (e.g., geographic names) into the *Walker-Matsumoto* combination.

267. As part of implementing its text classifier procedure, *Malsheen* teaches "generat[ing] eleven arrays 138 of classification values, each of which stores a single element ***for each 'word' in the paragraph being processed***. A word, for purposes of the enhanced text reader 128 is any chunk of text that requires separate processing. For instance, in strings having special punctuation such as slashes, hyphens, and colons, each text element separated by such punctuation is treated as a separate word. Thus the string '3/12/95' is treated as three words: 1) '3/', 2) '12/' and 3) '95'." Ex.1007 (*Malsheen*), 5:20-29.

268. In my opinion, person of ordinary skill in the art would have understood that for *Malsheen*'s text classifier procedure to generate the arrays "***for each word***" in the paragraph of text, *Malsheen*'s text classifier would have had to separate the text into its constituent words.

269. Indeed, *Malsheen* provides the example of the date 3/12/95 being separated into and treated as three words: 3/, 12/, and 95. In other words, it is my opinion that a person of ordinary skill in the art reading *Malsheen* would have understood *Malsheen* to teach that its text classifier procedure breaks the text input into its individual words, and then further processes each word to determine its classification.

270. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach breaking the raw text input into its constituent words ("tokenizing the text input").

(ii) 6[b]: automatically identifying the text normalization type of the plurality of tokens; and 6[c] automatically determining that the contrastive stress pattern is to be applied for the plurality of tokens.

271. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches Elements 6[b]-6[c] for the same reasons discussed in §§XI.A.4.a(i) (limitation 2[a]) - XI.A.4.a(ii) (limitation 2[b]), according to the table below.

Element	Section
6[b]	XI.A.4.a(i) (limitation 2 [a])
6[c]	XI.A.4.a(ii) (limitation 2[b])

272. For example, for the same reasons I discussed in §XI.A.4.a(i) (limitation 2[a]), a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach a computer identifying tokens "SC" and "NC" as abbreviations for state names ("automatically identifying the text normalization type of the plurality of tokens").

273. Further, as discussed in §§XI.A.1.c (limitation 1[b]) - XI.A.1.d (limitation 1[c]), *Matsumoto's* collation unit 14 compares the words in

corresponding "b" positions to determine how to assign variation coefficients for pitch, speed/duration, and volume to different words. Thus, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach automatically determining that a contrastive stress pattern is to be applied to the plurality of tokens (i.e., "NC" and "SC") ("automatically determining that the contrastive stress pattern is to be applied for the plurality of tokens.").

e. Claim 7: The method of claim 2, wherein the at least one token to be rendered with contrastive stress is identified based at least in part on an order of the plurality of tokens in the text input.

274. In my opinion, *Walker* in combination with *Matsumoto* and *Malsheen* teaches this claim.

275. As I described in §XI.A.1.c(v) (application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report), *Matsumoto* discloses that collation unit 14 collates two sets of a(n, b) such that words in the same "b" position are compared. The "b" position is determined by the order of words in a sentence. *See* Ex. 1006 (*Matsumoto*), ¶[0073]. Indeed, using the example weather report I have described herein, *Matsumoto*'s collation unit determines whether words in the second sentence (n=2)—the second occurrence of words in the same "b" position—are the same as, or different from, corresponding words in the first

sentence (n=1), and therefore, whether words in the second sentence (n=2) are assigned contrastive stress. *See* §XI.A.1.c(v) (application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report).

276. As I also described in §XI.A.1.c(v) (application of the *Walker-Matsumoto-Malsheen* combination to the example text weather report), a person of ordinary skill in the art would have understood that collation unit 14 compares corresponding portions of two tokens because "North" and "South" both appear in the b=5 position and "Carolina" and "Carolina" both appear in the b=6 position. Thus, in my opinion, a person of ordinary skill in the art would have understood *Matsumoto*'s collation unit 14 calculates the variation degree based on the order of a word within a sentence.

277. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach identifying words of a weather report to be the same or different between two or more sentences ("wherein the at least one token to be rendered with contrastive stress is identified") based on the order of how words appear in a sentence ("based at least in part on an order of the plurality of tokens in the text input").

f. Claims 9-11, 13

278. In my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach an apparatus performing the functions recited in claims 9-11 and 13 for the same reasons as described in §§XI.A.4.a (Claim 2) - XI.A.4.d (Claim 6), according to the table below.

Claim	Section
9	XI.A.4.a(i) (limitation 2[a]) and XI.A.4.a(ii) (limitation 2[b])
10	XI.A.4.b (Claim 3)
11	XI.A.4.c (Claim 4)
13	XI.A.4.d(i) (limitation 6[a]) and XI.A.4.d(ii) (limitation 6[b])

g. Claims 15-17 and 19-20

279. In my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach the at least one software module performing the functions recited in claims 15-17, and 19-20 for the same reasons as described in §§XI.A.4.a (Claim 2) - XI.A.4.e (Claim 7), according to the table below.

Claim	Section
15	XI.A.4.a(i) (limitation 2[a]) and XI.A.4.a(ii) (limitation 2[b])
16	XI.A.4.b (Claim 3)
17	XI.A.4.c (Claim 4)
19	XI.A.4.d(i) (limitation 6[a]) and XI.A.4.d(ii) (limitation 6[b])
20	XI.A.4.e (Claim 7)

A. Ground 1B: *Walker* in Combination With *Matsumoto* and *Bellegarda* Renders Obvious Claims 1-4, 6-11, 13-17, and 19-20

1. 1[c], 8[e], 14[c]

280. In my opinion, the *Walker-Matsumoto-Malsheen* combination teaches elements 1[c], 8[e], and 14[c] for the reasons I discussed in §§XI.A.1.d (limitation 1[c]), XI.A.2.c (limitation 8[e]), and XI.A.3.b (limitation 14[c]), respectively.

281. To the extent that Patent Owner argues and the Board agrees that deemphasizing a word/phrase constitutes assigning contrastive stress—and, therefore, *Walker-Matsumoto-Malsheen* does not meet the portions of elements 1[c], 8[e], and 14[c] that require not assigning contrastive stress to the identified at least one second portion of the at least one token—in my opinion, these claim limitations would nevertheless have been obvious to a person of ordinary skill in the art based on *Bellegarda*'s teachings to configure *Walker-Matsumoto-*

Malsheen's TTS engine such that words that are the same (e.g., "Carolina" at a(2, 6) as compared to "Carolina" at a(1, 6)) are not deemphasized.

282. For example, *Bellegarda* discloses a "speech synthesis system 100 [that] receives a text input 104" (*Bellegarda*, 5:8-9) and incorporates a "word prominence specification system 200" that "applies word prominence" (*id*., 5:21-26). *See also id*., FIG. 1. In one embodiment, *Bellegarda* discloses a speech synthesizer that stores and tracks previously spoken words in a buffer such that newly synthesized words that are the same as previous words are not emphasized and receive no change when spoken:

> For example, in one alternative embodiment, each new word is considered a candidate for prominence, and a list of previously spoken words is maintained in a FIFO (first-in-first-out) buffer having a specified depth. ***If a current word is already in the FIFO buffer, no accent is applied to the word when spoken***, but if the word is not in the buffer (i.e., the current word is a "new" word), prominence is applied to the word.

Id., 5:60-67.

283. Further, as I noted in §XI.A.1.d (limitation 1[c]), *Matsumoto*'s maximum and minimum variation coefficient values are user-settable. Thus, in my opinion, it would have been obvious for a person of ordinary skill in the art to set the minimum variation coefficient value to a value above 0, including 1.0.

284. As I discussed in §XI.A.1.d (limitation 1[c]), if the minimum variation coefficient is set to 1.0, there would be no change to the variation coefficients for words or phrases determined to be the same. Therefore, in my opinion, a person of ordinary skill in the art would have understood that the minimum variation coefficient value could be set to a value of 1.0.

285. Accordingly, in my opinion, a person of ordinary skill in the art would have found it obvious based on *Bellegarda*'s teaching to set the minimum variation coefficient to 1.0 so that there would be no change to the variation coefficients for words or phrases determined to be the same.

286. Indeed, only adjusting the prosody of words determined to be different between two sentences would have been obvious to a person of ordinary skill in the art because it would have amounted to the simple substitution of one known element (*Bellegarda* teaches not adjusting the prosody of corresponding words that are the same between two sentences) for another (*Matsumoto* teaches deemphasizing corresponding words that are the same between two sentences) to obtain predictable results (in the *Walker-Matsumoto-Malsheen-Bellegarda* combination corresponding words that different between two sentences would continue to receive contrastive stress).

287. Further, it is my opinion that a person of ordinary skill in the art would have also recognized that not adjusting the prosody of corresponding words that are the same between two sentences would have reduced the computations required by the *Walker-Matsumoto-Malsheen-Bellegarda* combinations inasmuch as the TTS engine would not have needed to adjust the speed/duration, pitch and volume coefficients of corresponding words that are the same between two sentences.

288. Thus, in my opinion, the combination would have also been obvious to a person of ordinary skill in the art because it amounted to use of a known technique (*Bellegarda* teaches not adjusting the prosody of words that are the same between two sentences) to improve similar devices (to reduce the computational requirements of *Walker-Matsumoto-Malsheen*'s TTS engine by) in the same way (by not adjusting the prosody of words that are the same between the two sentences in the *Walker-Matsumoto-Malsheen-Bellegarda* combination).

289. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen-Bellegarda* combination to teach assigning stress by increasing the speed/duration, pitch, and volume using variation coefficients C1, C2, and C3 to the identified first portion of the at least one token (e.g., "South" of "SC")—("assigning contrastive stress to the identified

at least one first portion of the at least one token") but not changing the speed/duration, pitch, and volume for the second portion of the at least one token (e.g., "Carolina" of "SC") —("but not to the identified at least one second portion of the at least one token").

2. All Other Features of Claims 1-4, 6-11, 13-17, and 19-20

290. In my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen-Bellegarda* combination to teach all other features of claims 1-4, 6-11, 13-17, and 19-20 for the reasons I previously discussed according to the table below.

Element/Claim	Section
1[pre]	XI.A.1.a (limitation 1[pre])
1[a]	XI.A.1.b (limitation 1[a])
1[b]	XI.A.1.c(i) through XI.A.1.c(v) (limitation 1[b])
1[d]	XI.A.1.e (limitation 1[d])
1[e]	XI.A.1.f (limitation 1[e])
2[a]	XI.A.4.a(i) (limitation 2[a])
2[b]	XI.A.4.a(ii) (limitation 2[b])
3	XI.A.4.b (Claim 3)
4	XI.A.4.c (Claim 4)

Element/Claim	Section
6	XI.A.4.d(i) (limitation 6[a]) and XI.A.4.d(ii) (limitation 6[b])
7	XI.A.4.e (Claim 7)
8[pre]	XI.A.2.a (limitation 1[pre]
8[a]	XI.A.2.b (limitation 8[a])
8[b]	XI.A.2.b (limitation 8[b]])
8[c]	XI.A.1.b (limitation 1[a])
8[d]	XI.A.1.c(i) through XI.A.1.c(v) (limitation 1[b])
8[f]	XI.A.1.e (limitation 1[d])
8[g]	XI.A.1.f (limitation 1[e])
9	XI.A.4.a(i) (limitation 2[a]) and XI.A.4.a(ii) (limitation 2[b])
10	XI.A.4.b (Claim 3)
11	XI.A.4.c (Claim 4)
13	XI.A.4.d(i) (limitation 6[a]) and XI.A.4.d(ii) (limitation 6[b])
14[pre]	XI.A.3.a (limitation 14[pre])
14[a]	XI.A.1.b (limitation 1[a])
14[b]	XI.A.1.c(i) through XI.A.1.c(v) (limitation 1[b])

Element/Claim	Section
14[d]	XI.A.1.e (limitation 1[d])
14[e]	XI.A.1.f (limitation 1[e])
15	XI.A.4.a(i) (limitation 2[a]) and XI.A.4.a(ii) (limitation 2[b])
16	XI.A.4.b (Claim 3)
17	XI.A.4.c (Claim 4)
19	XI.A.4.d(i) (limitation 6[a]) and XI.A.4.d(ii) (limitation 6[b])
20	XI.A.4.e (Claim 7)

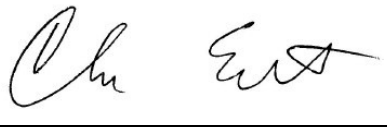
XII. SECONDARY CONSIDERATIONS

291. I am not aware of any evidence in the '291 patent's prosecution history or elsewhere supporting any secondary considerations arguments, or evidence of nexus of such alleged evidence to the challenged claims. *See generally* Ex. 1004 (*'291FH*). To the extent Patent Owner asserts the existence of any secondary considerations in its responses, I reserve the right to address any such evidence.

292. I declare that all statements made herein of my knowledge are true, that all statements made on information and belief are believed to be true, and that these statements were made with the knowledge that willful false statements and

the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Executed on: January 21, 2025



Christopher M. Schmandt