

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 IPR2025-00459

U.S. Patent No. 8,825,486

**DECLARATION OF CHRISTOPHER M. SCHMANDT IN SUPPORT
OF PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 8,825,486**

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	BACKGROUND AND QUALIFICATIONS	2
III.	MATERIALS REVIEWED	7
IV.	LEVEL OF ORDINARY SKILL IN THE ART	9
V.	RELEVANT LEGAL STANDARDS	11
VI.	SUMMARY OF OPINIONS	15
VII.	OVERVIEW OF THE '486 PATENT	15
VIII.	OVERVIEW OF THE PRIMARY PRIOR ART REFERENCES	20
	A. Analogous Art	20
	B. <i>Walker</i>	22
	C. <i>Matsumoto</i>	26
IX.	CLAIM CONSTRUCTION	31
X.	SPECIFIC GROUNDS FOR CHALLENGE	32
XI.	DETAILED EXPLANATION OF GROUNDS	33
	A. Ground 1A: <i>Walker</i> in Combination With <i>Matsumoto</i> Renders Obvious Claims 1, 7-8, and 14-15	33
	1. Motivation to Combine <i>Walker</i> and <i>Matsumoto</i>	33
	2. Independent Claim 1	40
	a. 1[pre]: A method for use with a speech-enabled application, the method comprising:	40

b.	1[a]: receiving, from the speech-enabled application, input comprising a plurality of text strings;	44
c.	1[b]: identifying a first portion of a first text string of the plurality of text strings as differing from a corresponding first portion of a second text string of the plurality of text strings, and a second portion of the first text string as not differing from a corresponding second portion of the second text string;	48
d.	1[c]: assigning contrastive stress to the first portion of the first text string and/or to the corresponding first portion of the second text string, but not to the second portion of the first text string, and not to the corresponding second portion of the second text string;	60
e.	1[d]: generating, using at least one computer system, speech synthesis output to render the plurality of text strings as speech having the assigned contrastive stress; and.....	76
f.	1[e]: providing the speech synthesis output for the speech-enabled application.....	79
3.	Independent Claim 8	80
a.	8[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 use with a speech-enabled application, the method comprising:	80
b.	8[a]-8[e]	84
4.	Independent Claim 15	85
a.	15[Pre]: A method for generating speech output via a speech-enabled application, the method comprising.....	85

b.	15[a]: generating, using at least one computer system executing the speech-enabled application, a plurality of text strings, each of the plurality of text strings corresponding to a portion of a desired speech output;..86	
c.	15[b]: inputting the plurality of text strings to at least one software module configured to identify a first portion of a first text string of the plurality of text strings as differing from a corresponding first portion of a second text string of the plurality of text strings, and a second portion of the first text string as not differing from a corresponding second portion of the second text string; 90	
d.	15[c]: receiving, from the at least one software module, speech synthesis output to render the plurality of text strings with contrastive stress assigned to the first portion of the first text string and/or to the corresponding first portion of the second text string, but not to the second portion of the first text string, and not to the corresponding second portion of the second text string; and.....94	
e.	15[d]: generating, using the speech synthesis output, an audio speech output corresponding to the desired speech output. 96	
5.	Dependent Claims 7 and 14 99	
a.	Claim 7: The method of claim 1, wherein the speech synthesis output comprises an indication of the first portion of the first text string and/or the corresponding first portion of the second text string as being assigned contrastive stress.....99	
b.	Claim 14: The at least one non-transitory computer-readable storage medium of claim 8, wherein the speech synthesis output comprises an indication of the first portion of the first text string and/or the corresponding	

	first portion of the second text string as being assigned contrastive stress.....	101
B.	Ground 2A: <i>Walker</i> in Combination With <i>Matsumoto</i> and <i>Malsheen</i> Renders Obvious Claims 2-5, 9-12, 16-19.....	102
1.	Claim 2: The method of claim 1, wherein the identifying comprises identifying the first portion of the first text string as differing from the corresponding first portion of the second text string based at least in part on a normalized orthography of the first and second text strings.....	102
2.	Claim 3: The method of claim 1, wherein the first and second text strings represent different numerical fields within a larger text string.....	108
3.	Claim 4: The method of claim 3, wherein the numerical fields are selected from the group consisting of: currency fields, date fields, digit sequence fields, number fields, fractional number fields, ordinal number fields, telephone number fields, flight number fields, street number fields, time fields, and zipcode fields.....	110
4.	Claim 5: The method of claim 1, wherein the receiving comprises receiving the first and second text strings as first and second parameters passed to a function called by the speech- enabled application to render the first and second text strings with a contrastive stress pattern.	110
5.	Claims 9-12.....	115
6.	Claims 16-19.....	116
a.	Claims 16-18.....	117
b.	Claim 19.....	117
C.	Ground 3A: <i>Walker</i> in Combination With <i>Matsumoto</i> and <i>Marple</i> Renders Obvious Claims 6, 13, and 20.....	120

1.	Claim 6: The method of claim 1, wherein the speech synthesis output comprises identification of a plurality of audio recordings to render the plurality of text strings as speech, at least one of the plurality of audio recordings being selected to render the first portion of the first text string and/or the first portion of the second text string as speech carrying contrastive stress.....	120
2.	Claim 13: The at least one non-transitory computer-readable storage medium of claim 8, wherein the speech synthesis output comprises identification of a plurality of audio recordings to render the plurality of text strings as speech, at least one of the plurality of audio recordings being selected to render the first portion of the first text string and/or the first portion of the second text string as speech carrying contrastive stress.	133
3.	Claim 20: The method of claim 15, wherein the speech synthesis output comprises identification of a plurality of audio recordings to render the plurality of text strings as speech, at least one of the plurality of audio recordings being selected to render the first portion of the first text string and/or the first portion of the second text string as speech carrying contrastive stress.....	134
D.	Ground 1B: <i>Walker</i> in Combination With <i>Matsumoto</i> and <i>Bellegarda</i> Renders Obvious Claims 1, 7-8, 14-15	134
1.	1[c], 8[c], 15[c]	134
2.	All Other Features of Claims 1, 7, 8, 14, and 15	138
E.	Ground 2B (<i>Walker-Matsumoto-Bellegarda-Malsheen</i>)	139
F.	Ground 3B (<i>Walker-Matsumoto-Bellegarda-Marple</i>)	140
XII.	SECONDARY CONSIDERATIONS	141

I. INTRODUCTION

1. I have been retained by Samsung Electronics Co., Ltd. and Samsung Electronics America, Inc. (collectively “Petitioners” or “Samsung”) 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 1-20 of U.S. Patent No. 8,825,486 (“the ’486 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 ’486 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 '486 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 8,825,486 (“’486*Pat*”) (Ex. 1001);
- File History of U.S. Patent 8,825,486 (“’486*FH*”) (“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);
- U.S. Patent App. No. 2008/0195391 (“*Marple*”) (Ex. 1008);
- U.S. Patent 7,010,489 (“*Lewis*”) (Ex. 1009);
- U.S. Patent 7,177,807 (“*Schmid*”) (Ex. 1010);
- U.S. Patent 8,290,775 (“*Etezadi*”) (Ex. 1011);
- U.S. Patent 6,829,577 (“*Gleason*”) (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);
- Excerpts from MICROSOFT COMPUTER DICTIONARY (5th ed. 2002) (“*Computer Dictionary*”) (Ex. 1016);
- U.S. Patent 7,779,357 (“*Naik*”) (Ex. 1017);
- U.S. Patent 7,937,125 (“*May*”) (Ex. 1018);
- U.S. Patent 8,447,610 (“’610Pat”) (Ex. 1019);
- Summary of all patents/applications in the ’486 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 ’486 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 '486 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 1-20 ("Challenged Claims") of the '486 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 '486 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 '486 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 '486 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 1-20 of the '486 patent are rendered obvious by the prior art.

VII. OVERVIEW OF THE '486 PATENT

35. The '486 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 ('486Pat), Abstract.

36. The '486 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., 8:25-28.

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., 8:37-53.³

37. For example, the '486 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.*, 12:41-47.

38. The '486 patent achieves this effect by first receiving “a text transcription of a desired speech output.” *Id.*, 4:49-50. A synthesis system

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

“analyze[s] the text input to determine which portion(s) of the speech output should carry contrastive stress.” *Id.*, 9:11-13. “[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.*, 9:26-31.

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.*, 4:51-52. 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.*, 4:52-54.

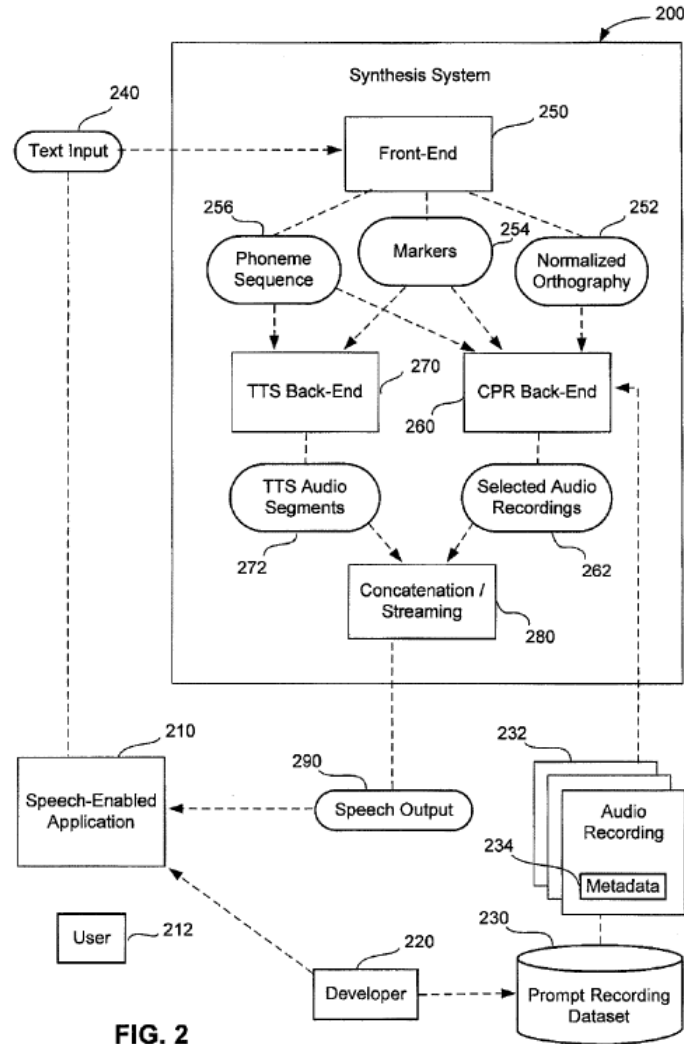


FIG. 2

Ex. 1001 ('486Pat), 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

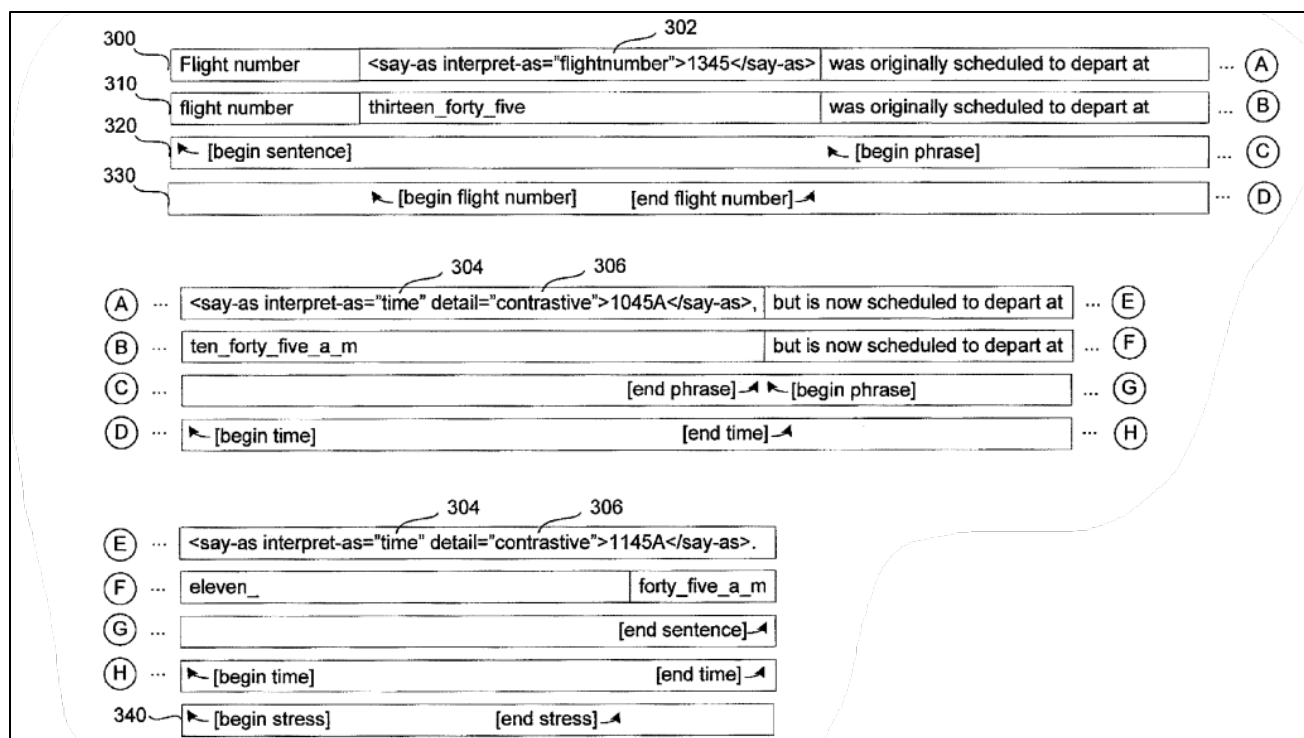
‘11:45 a.m.’ that differs from and contrasts with the ‘ten’ of ‘10:45 a.m.’ (i.e., the ‘forty_five’ and the ‘a.m.’ do not differ or contrast).” Ex. 1001 (’486Pat), 23:4-12.

41. The ’486 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 (’486Pat), 15:6-18.

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.*, 24:11-24. 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.*, 24:25-29; *See also id.*, 16:25-47.



Ex. 1001 ('486Pat), 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 '486 patent because they are within the same field of endeavor and reasonably pertinent to at least one problem addressed by the '486 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 '486 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 '486 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 '486 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. *Marple* discloses a “speech synthesizer for synthesizing speech from text.” Ex. 1008 (*Marple*), ¶[0013]. Like the ’486 patent, *Marple* is concerned with improving “synthetic, or ‘machine’ speech to ‘humanize’ it to sound more appealing and natural to the human ear.” *Id.*, ¶[0037]. *Marple* discloses using a database of pre-recorded phonemes which are assembled based on the corresponding phonemes of input text. *See id.*, ¶¶[0013], [0050]-[0051].

48. *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 ’486 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.

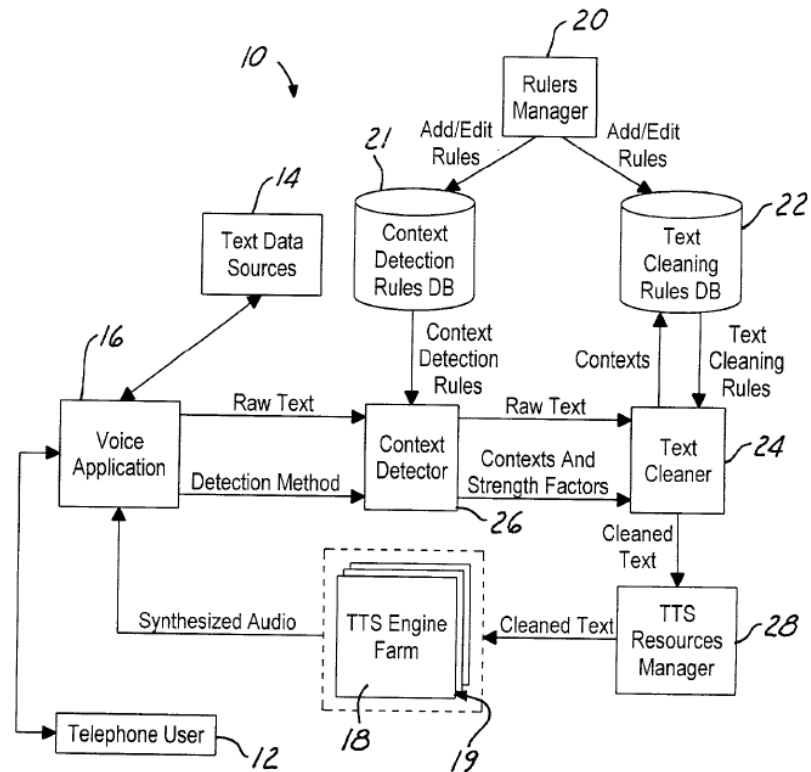
B. *Walker*

49. *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].

50. *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]).

51. 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.

52. *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].

53. 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].

54. *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].

55. 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]).

56. 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].

57. 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*

58. *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].

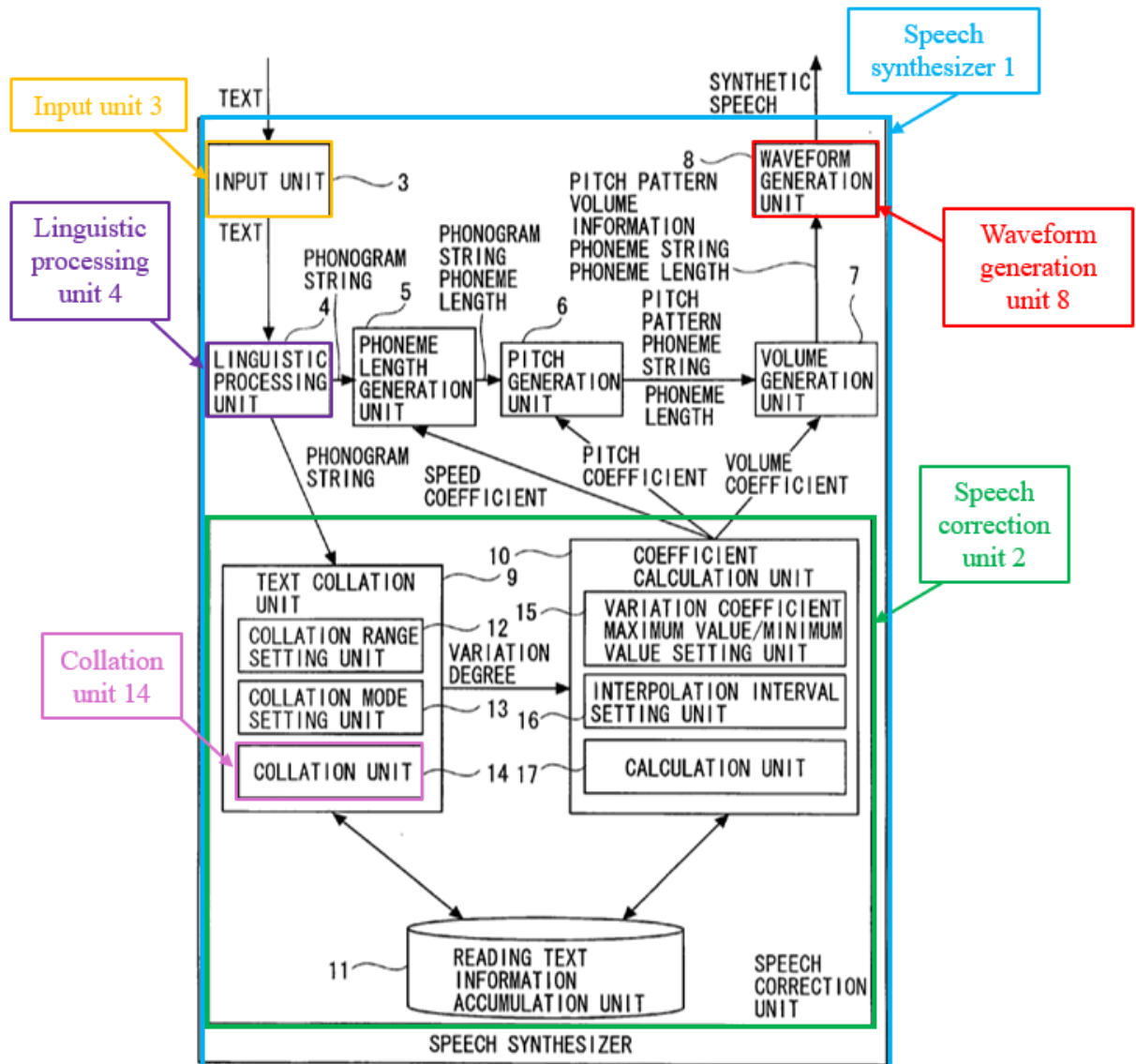
59. However, if “the speech synthesizer vocalizes such sentences in a monotone ... the hearer can not [sic] concentrate on a want-to-hear point in the speech and might fail to hear the want-to-hear point.” *Id.*

60. 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.*).

61. 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].

62. 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).

63. “[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]).

64. 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].

65. 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].

66. 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.

67. 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.*

68. 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].

69. *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

70. I have given all the claim terms their plain 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

71. In the sections below, I explain in detail how claims 1-20 of the '486 patent are not patentable over the prior art. More particularly:

72. **Ground 1A** challenges claims 1, 7-8, and 14-15 as obvious over the combination of *Walker* and *Matsumoto*.

73. **Ground 1B** challenges claims 1, 7-8, and 14-15 as obvious over the combination of *Walker*, *Matsumoto*, and *Bellegarda*.

74. **Ground 2A** challenges claims 2-5, 9-12, and 16-19 as obvious over the combination of *Walker*, *Matsumoto*, and *Malsheen*.

75. **Ground 2B** challenges claims 2-5, 9-12, and 16-19 as obvious over the combination of *Walker*, *Matsumoto*, *Bellegarda*, and *Malsheen*.

76. **Ground 3A** challenges claims 6, 13, and 20 as obvious over the combination of *Walker*, *Matsumoto*, and *Marple*.

77. **Ground 3B** challenges claims 6, 13, and 20 as obvious over the combination of *Walker*, *Matsumoto*, *Bellegarda*, and *Marple*.⁴

XI. DETAILED EXPLANATION OF GROUNDS

A. Ground 1A: *Walker* in Combination With *Matsumoto* Renders Obvious Claims 1, 7-8, and 14-15

1. Motivation to Combine *Walker* and *Matsumoto*

78. 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.

79. *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

⁴ To the extent my declaration refers to prior art references other than those recited specifically in each of the above-listed grounds 1A, 1B, 2A, 2B, 3A, and 3B, 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.

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].

80. *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]).

81. *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].

82. 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*.

83. *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].

84. 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].

85. 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.

86. 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.

87. 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.

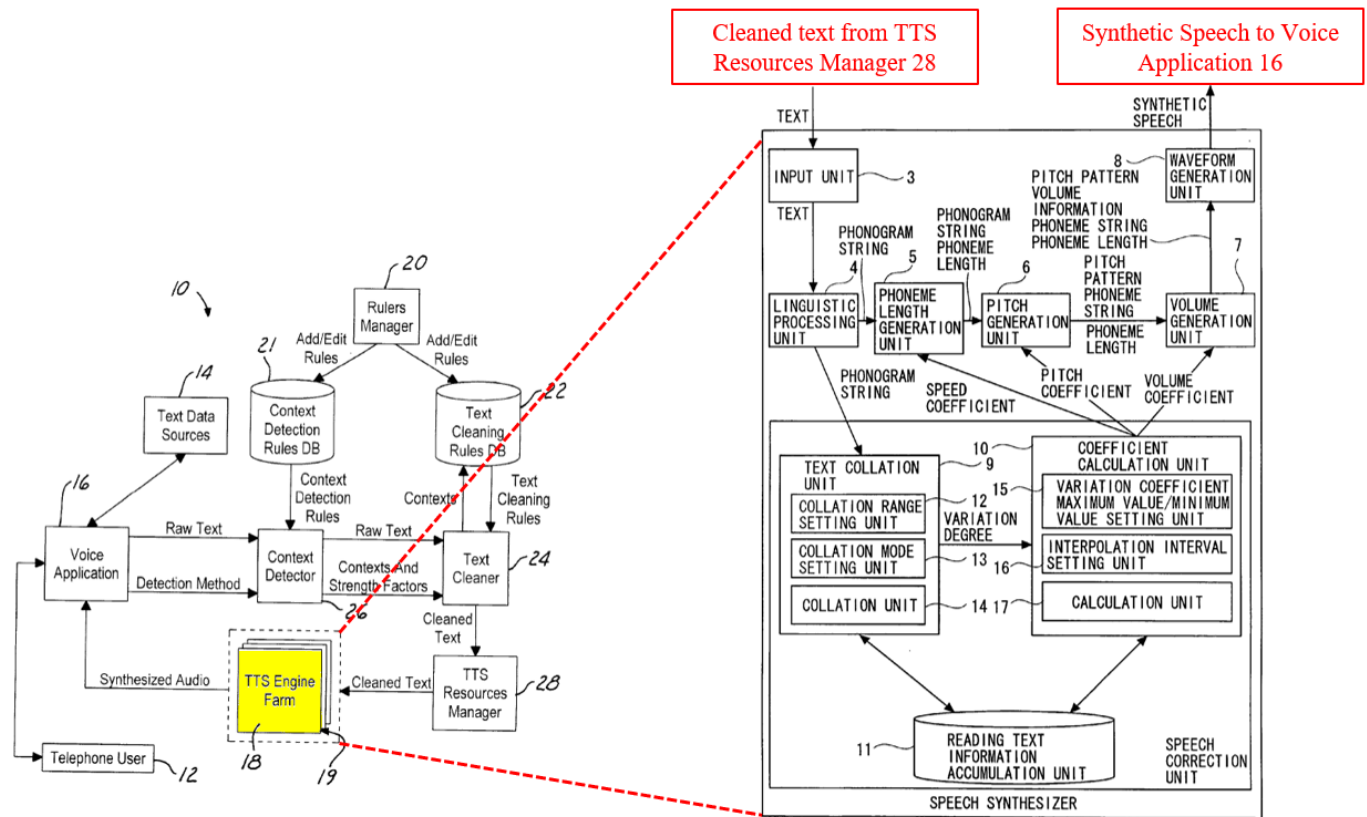
88. In my opinion, a person of ordinary skill in the art would have appreciated that *Matsumoto*’s speech synthesizer would have advanced *Walker*’s objective of providing more natural synthesized speech for a weather report and provided additional details regarding how to implement such a TTS engine.

89. 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.

90. 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)).

91. 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.



92. 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).

93. 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].

94. 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.

2. Independent Claim 1

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

95. I understand that "[a] method for use with a speech-enabled application, the method comprising" is the preamble to claim 1 of the '486 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* teaches the preamble.⁵

96. *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].

97. 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].

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.

98. 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.”).

99. *Walker*'s voice application 16 is consistent with how the '486 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 ('486*Pat*), 1:30-32.

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., 1:35-57.

100. The '486 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., 9:39-63.

101. Accordingly, in my opinion, 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 use with a speech-enabled application").

b. 1[a]: receiving, from the speech-enabled application, input comprising a plurality of text strings;

102. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.

103. *Walker* discloses that context detector 26 receives raw text (i.e., unprocessed text) from voice application 16. *See* 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.2.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, in my opinion, a person of ordinary skill in the art would have understood context detector 26 receives raw text from a speech-enabled application 16.

104. *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].

105. 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.”).

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

107. 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 string(s) that pertain 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.

108. 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).

109. 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).

110. 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 strings), each describing the weather for a different location/city or day, as taught by *Matsumoto*.

111. Further, it is my opinion that a person of ordinary skill in the art would have understood a “text string” to be a broad phrase that encompasses any combination of word(s), phrase(s), and/or sentence(s).

112. For example, U.S. Pat. No. 8,290,775 (“*Etezadi*”) discloses a method for pronunciation correction between different languages in text-to-speech (TTS) systems and speech recognition (SR) systems. *See* Ex. 1011 (*Etezadi*), Abstract. *Etezadi* states that “in order to generate a phoneme-based text string that may be recognized by a target language-based TTS and/or speech recognition system, a mapping of the phonemes comprising the starting language word/phrase is performed to corresponding phonemes of any ending or target language.” *Id.*, 7:19-24. *Etezadi* provides examples of text strings that are synthesized into speech, including the “English language text string ‘The Beatles’[.]” *Id.*, 1:26-28.

113. In addition, U.S. Patent No. 6,829,577 (“*Gleason*”) discloses a method for synthesizing vowels in a speech synthesizer. *See* Ex. 1012 (*Gleason*), 1:16-18. *Gleason* provides examples of input text that are processed into speech, including “text string ‘32’” and “text string ‘=’.” *Id.*, 3:65-67.

114. This understanding is consistent with the '486 patent which describes an entire sentence as a “text string”:

Text input 300 is an exemplary text string that speech-enabled application 210 may generate and submit to synthesis system 200, to request that synthesis system 200 provide a synthesized speech output rendering this desired speech output as audio speech.

Ex. 1001 (*'486Pat*), 19:36-40; *see also id.*, FIG. 3A.

115. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* combination to teach a context detector receiving from a voice application (“receiving, from the speech-enabled application”) the raw text of a weather report including two or more sentences (“input comprising a plurality of text strings”).

c. 1[b]: identifying a first portion of a first text string of the plurality of text strings as differing from a corresponding first portion of a second text string of the plurality of text strings, and a second portion of the first text string as not differing from a corresponding second portion of the second text string;

116. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.

117. For the reasons I discussed in §XI.A.2.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.

	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.

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]

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

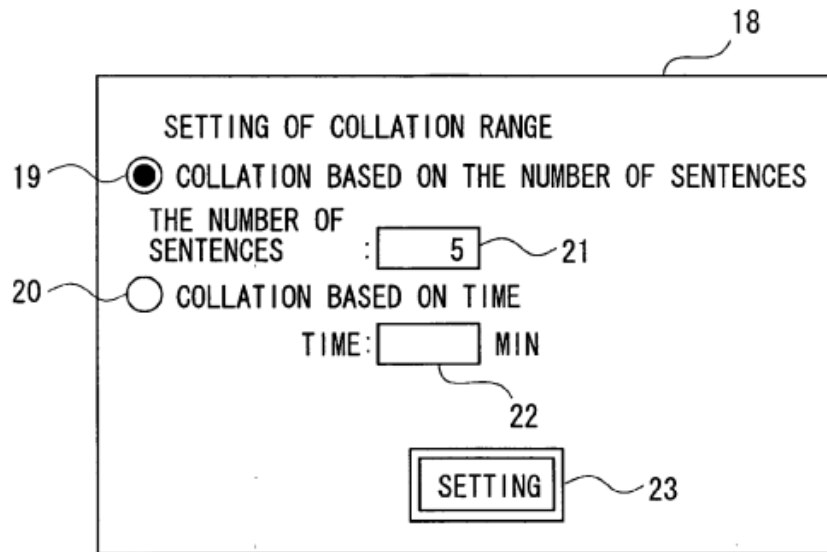
Ex. 1006 (*Matsumoto*), Expression 1.

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.

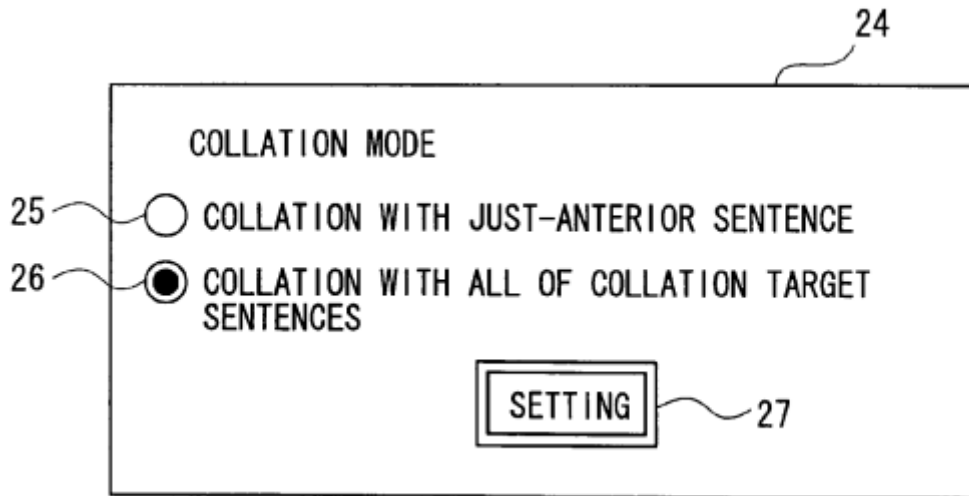
123. Indeed, *Matsumoto*'s use of radio buttons in figure 2 would have conveyed to a person of ordinary skill in the art that collation based on a number of sentences is one of two alternative choices (i.e., the other choice being a collation based on time). Thus, in my opinion, a person of ordinary 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. In the discussion that follows, I assume that the weather report received by *Walker*'s voice application 16 from text data source 14 is comprised of

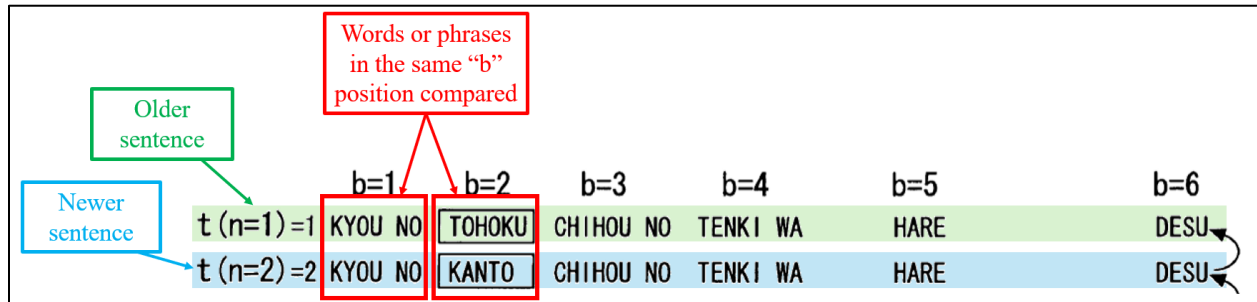
only the first and second sentences in *Matsumoto*'s figure 6, and that *Matsumoto*'s TTS system is configured to perform collation based on the number of sentences (i.e., two sentences) and to collate the latest sentence with only the anterior sentence. In the example weather report discussed in this ground, the second sentence (i.e., $t(n=2)$) corresponds to "a first text string of the plurality of text strings," and the first sentence (i.e., $t(n=1)$) corresponds to "a second text string of the plurality of text strings."

127. As I discussed in §XI.A.1 (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.

128. *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. See Ex. 1006 (*Matsumoto*), ¶¶[0044], [0074].

129. 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].



Ex. 1006 (*Matsumoto*), FIG. 6 (excerpted and annotated).

130. As *Matsumoto* discusses with reference to figure 6, “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.” *Id.*, ¶[0074].

131. *Matsumoto* further discloses that when corresponding words or phrases between two sentences are determined to be the same (e.g., “kyou no” and “kyou no”) the variation degree is “0,” but if determined to be different (e.g., “Tohoku” and “Kanto”), the variation degree is “1.” *See id.*, ¶[0080]; Expression 1.

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

[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.

133. 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].

134. When m=1, the expression $\delta(a(m, b), a(m-1, b))$ is evaluated as $\delta(a(1, b), a(1-1, b))$, i.e., $\delta(a(1, b), a(0, b))$. As *Matsumoto* notes, “a(0, b) = a(1, b).” *Id.*, ¶[0080]. Thus, when m=1, $\delta(a(m, b), a(m-1, b))$ in Expression 1 is “1.” *See id.*, ¶[0080] (“[I]n the mathematical expression (1), $\delta(a(m, b), a(m-1, b))$ 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.

135. When m=2, the expression $\delta(a(m, b), a(m-1, b))$ is evaluated as $\delta(a(2, b), a(2-1, b))$, i.e., $\delta(a(2, b), a(1, b))$. In this case, the second sentence (n=2) is being compared to the first sentence (n=1). When calculating the variation degree for “Kanto” (i.e., a(2, 2)), b=2 because “Kanto” appears as the second word in the sentence. Thus, when “Kanto” (a(2, 2)) is compared to “Tohoku” (a(1, 2)), $\delta(a(m,$

b), $a(m-1, b)$) in Expression 1 is “0” because $a(2, 2)$ is not equal to $a(1, 2)$. *See id.*, ¶[0080] (“Further, in the mathematical expression (1), $\delta(a(m, b), a(m-1, b))$... represents ‘0’ when $a(m, b)$ is not equal to $a(m-1, b)$.”).

136. Second, the variable $t(n)$ represents the time or order of the input:⁶

[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].

137. Thus, “ n ” in Expression 1 is “2” because the second sentence ($n=2$) is being accumulated.

138. 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

⁶ 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].

second input (i.e., the second sentence (n=2)) and t(1) represents the first input (i.e., the first sentence (n=1)).

139. Accordingly, as shown below using Expression 1, the variation degree for “Kanto” (a(2, 2)) is calculated to be “1.”

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

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

140. For the same reasons I discussed above, the variation degree for “kyou no” (a(2, 1)) is calculated to be “0” because “kyou no” (a(2, 1)) is the same as “kyou no” (a(1, 1)). More particularly, unlike the variation degree calculation for “Kanto” described above, in this calculation, $\delta(a(m, b), a(m-1, b))$ is “1” because a(2, 1) is equal to a(1, 1). See Ex. 1006 (*Matsumoto*), ¶[0080] (“[I]n the mathematical expression (1), $\delta(a(m, b), a(m-1, b))$ represents ‘1’ when a(m, b) is equal to a(m-1, b) ...”). Therefore:

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

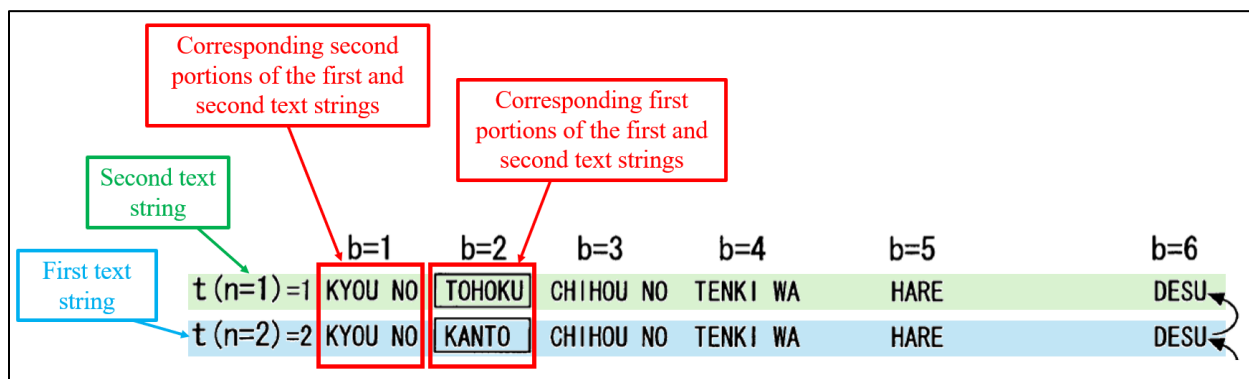
$$\frac{1-1}{2-1+1} + \frac{1-1}{2-2+1} = 0$$

141. Similar to “kyou no” as described above, the calculated variation degree for each of the remaining words or phrases in the second sentence (n=2)

(i.e., $a(2, 3)$, $a(2, 4)$, $a(2, 5)$ and $a(2, 6)$) are also calculated to be “0” because each word or phrase is determined to be the same as the corresponding word or phrase in the first sentence $n=1$. See also *id.*, FIG. 6.

142. 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 two text strings because “Kanto” and “Tohoku” both appear in the $b=2$ position and “kyou no” and “kyou no” both appear in the $b=1$ position.

143. In my opinion, a person of ordinary skill in the art would have also understood that ***portions*** of two text strings are compared because “Kanto,” “Tohoku,” “kyou no,” and “kyou no” are all words or phrases within a larger sentence of a weather report: $t(n=1)$ kyou no Tohoku chihou no tenki wa hare desu ([Today’s weather in the Tohoku region is fair.]), and $t(n=2)$ kyou no Kanto chihou no tenki wa kumori desu ([Today’s weather in the Kanto region is cloudy]).



Ex. 1006 (*Matsumoto*), FIG. 6 (excerpted and annotated).

144. Additionally, it is my opinion that a person of ordinary skill in the art would have understood that in the scenario discussed herein, “Kanto” appearing in location a(2, 2) corresponds to “a first portion of a first text string” and “kyou no” (today’s) appearing in location a(2, 1) corresponds to “a second portion of the first text string.”

145. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* combination to teach comparing each word/phrase in a current sentence of the weather report to a word/phrase in the same “b” position in a prior sentence, and calculating a variation degree, such as “1” if the words/phrases are different (“identifying a first portion of a first text string of the plurality of text strings as differing from a corresponding first portion of a second text string of the plurality of text strings”), or “0” if there is no difference between a word/phrase in the current sentence and a word/phrase in the same “b” position in the prior sentence (“and a second portion of the first text string as not differing from a corresponding second portion of the second text string”).

d. 1[c]: assigning contrastive stress to the first portion of the first text string and/or to the corresponding first portion of the second text string, but not to the second portion of the first text string, and not to the corresponding second portion of the second text string;

146. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.⁷

147. As I discussed in §XI.A.2.c (limitation 1[b]), in the example weather report discussed in this ground, “Kanto” in location a(2, 1) corresponds to “a first portion of a first text string,” and “kyou no” (today’s) in location a(2, 2) corresponds to “a second portion of the first text string.”

148. *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

⁷ I understand that to show that the feature “assigning contrastive stress to the first portion of the first text string *and/or* to the corresponding first portion of the second text string” is taught by the prior art, it is sufficient to show that the prior art teaches *either* “assigning contrastive stress to the first portion of the first text string” *or* “assigning contrastive stress ... to the corresponding first portion of the second text string.”

corresponding words or phrases of two or more sentences. Ex. 1006 (*Matsumoto*), ¶[0054].

149. 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 multiplies the normal volume by a ***volume coefficient*** obtained from the coefficient calculation unit 10 (the calculation unit 17), thereby correcting the volume ...”).

150. 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.”).

151. 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.

152. 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

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

154. 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.

155. Thus, in my opinion, a person of skill in the art would have understood that prosodic characteristics are used to convey meaning in speech

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 Grounds 1B, 2B, and 3B that I discuss later in §§XI.D, XI.E, and XI.F, respectively, *Bellegarda* is **not** part of Grounds 1A, 2A and 3A.

because such characteristics are used to make speech sound more natural/realistic and therefore, more intelligible.

156. *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.”).

157. 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].

158. *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].

159. *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].

160. 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.

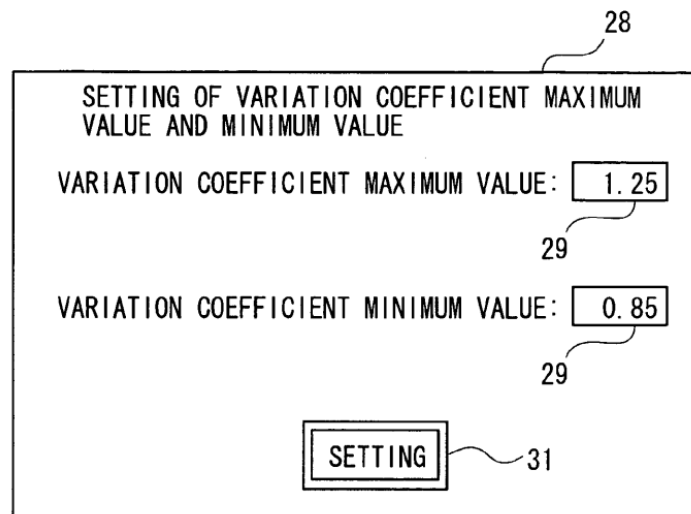
Id., ¶[0091].

161. 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.

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



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

163. 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.

164. 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 multiplies 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.

165. 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.”).

166. To calculate the volume variation coefficient (C3) for “Kanto,” *Matsumoto* discloses using Expression 7, shown below. *See id.*, ¶[0087] (“the

volume coefficient is calculated by employing the following mathematical expression (7).”⁹

[Mathematical Expression 7]

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

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

167. First, as I previously described in §XI.A.2.c (limitation 1[b]), the variation degree for “Kanto” (i.e., a(2, 2)) is calculated to be “1” because “Kanto” is different from “Tohoku” (i.e., a (1, 2)). Thus, v(n, b) =1.

168. Second, as I discussed in §XI.A.2.c (limitation 1[b]), 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, $\sum_{b=1} v(n, b) = 1$.

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.

169. 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 “Kanto,” the normal phoneme length is 5 because “Kanto” is made up of 5 phonemes (i.e., $f=5$). Similarly, the normal sentence length is 32 because the second sentence ($n=2$) comprises 32 phonemes (i.e., $g=32$). Thus, in Expression 7, $g=32$ and $f=5$.¹⁰

170. Fourth, *Matsumoto* discloses that $e(\text{MIN})$ is “a preset coefficient minimum value $e(\text{MIN})$.” *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 $e(\text{MIN})$ 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

¹⁰ My analysis of *Matsumoto*’s sentences and equations is based on the English translation of these sentences as they appear in *Matsumoto*. The analysis, however, would also be applicable to the sentences in Japanese after accounting for differences in the phonemes used in the two languages.

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.

171. Using each of the values I described above, C3 for “Kanto” (a(2, 2)) is calculated to be 4.8 as shown below.

$$C3(n, b) = \frac{v(n,b)ge(MIN)}{f \sum_{b=1} v(n,b)} = \frac{(1)(32)(0.75)}{(5)(1)} = 4.8$$

172. 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 4.8, 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.”).

173. Accordingly, the variation coefficient for volume (C3) for “Kanto” is set to 1.25. As such, the volume for “Kanto” would be increased by 25% when synthesized into speech.

174. 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.”

175. 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 “kyou no” ($a(2, 1)$), is calculated to be “0” as shown below.

$$C3(n, b) = \frac{v(n, b)ge(MIN)}{f \sum_{b=1} v(n, b)} = \frac{(0)(32)(0.75)}{(4)(1)} = 0$$

176. 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].

177. 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.).

178. Thus, in the specific example in which only *Matsumoto*’s first two sentences are collated (i.e., $n=2$ and $n=1$), the volume variation coefficient for “Kanto” ($a(2, 2)$) is 1.25 while the volume variation coefficient for “kyou no” ($a(2, 1)$) is 0.85. Thus, when corrected, the volume for “Kanto” is increased, while the volume for “kyou no” is decreased. Accordingly, only the word determined to be different between two text strings (i.e., “Kanto”) is emphasized while words or phrases that are the same (e.g., “kyou no”) are deemphasized. Additionally, in my opinion, a person of ordinary skill in the art would have understood *Matsumoto* to teach that, when collation is based on the number of sentences and comparing the latest sentence with the anterior sentence, and there are two sentences, only the words or phrase in the new sentence are assigned contrastive stress. See *Matsumoto*, ¶¶[0077]-[0080], [0105]. Thus, in my opinion, a person of ordinary skill in the art would have understood that “kyou no” in the first sentence (i.e., $a(1,$

1)) would not have been emphasized/ deemphasized (“not [assigning contrastive stress] to the corresponding second portion of the second text string”).

179. 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 second portion of the first text string,” as recited in claim limitation 1[c] and understood in the context of the ’486 patent.

180. The ’486 patent describes “contrastive stress” as *increasing* the prosody for the words/phrases meant to contrast with each other. *See* Ex. 1001 (’486*Pat*), 9:28-31 (“[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.”), 12:47-51 (“[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.”).

181. The ’486 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.

182. Additionally, during prosecution of U.S. Patent 8,447,610 (“the ’610 patent”)—which is the grand parent application of the ’486 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.” Ex. 1013 (’610FH), 83.

183. 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. 1019 (’610Pat), 8:12-15; Ex. 1001 (’486Pat), 8:21-25. 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. 1019 (’610Pat), 8:15-18; Ex.1001 (’486Pat), 8:25-28.

184. 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:

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.

185. Accordingly, in my opinion, both the specification and prosecution history of the '486 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.

186. 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.

187. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* 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 first text string—e.g., a(2, 2) ("Kanto")—"assigning contrastive stress to the first portion of the first text string ...") but not increasing the speed/duration, pitch, and volume for the second portion of the first text string—e.g., a(2, 1) ("kyou no")—"but not to the second portion of the first text string") and not increasing the speed/duration, pitch, and

volume for the second portion of the second text string—e.g., a(1, 1) (“kyou no”) (“and not to the corresponding second portion of the second text string”).

- e. **1[d]: generating, using at least one computer system, speech synthesis output to render the plurality of text strings as speech having the assigned contrastive stress; and**

188. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.

189. *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 ***multiplies the normal volume by a volume coefficient*** obtained from the coefficient calculation unit 10 (the calculation unit 17), thereby correcting the volume ...

Id., ¶[0105].

190. 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.”).

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

192. 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**.”).

193. 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].

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

195. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* combination to teach using a computer system to generate synthesized speech output of a text weather report (“generating, using at least one computer system, speech synthesis output”) with changes to speed/duration, pitch and volume using the assigned coefficients C1, C2, and C3 (“to render the plurality of text strings as speech having the assigned contrastive stress”).

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

196. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.

197. *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].

198. As I discussed in §XI.A.1 (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.

199. As I discussed in §XI.A.2.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.

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

3. Independent Claim 8

a. 8[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 use with a speech-enabled application, the method comprising:

201. 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 use with a speech-enabled application, the method comprising” is the preamble to claim 8 of the ’486 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* teaches the preamble.

202. For the reasons I discussed in §XI.A.2.a (limitation 1[pre]), *Walker* teaches “a method for use with a speech-enabled application.”

203. *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.”).

204. 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.

205. 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 (’000*Prov.*), 5.

206. 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

207. 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 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).

208. 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

¹¹ *Bellegarda* is not part of Ground 1A. See footnote **Error! Bookmark not defined.**

device, it is my opinion that this feature would have been obvious based on *Walker* in combination with *Matsumoto*.

209. 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* combination as computer-executable instructions stored on a non-transitory computer-readable medium because *Matsumoto* expressly teaches to do so.

210. 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* combination to store the program instruction executed by the computer).

211. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* 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. 8[a]-8[e]

212. In my opinion, *Walker* in combination with *Matsumoto* teaches Elements 8[a]-8[e] for the same reasons I discussed in §§XI.A.2.b (limitation 1[a]) - XI.A.2.f (limitation 1[e]), according to the table below.

Element	Section
8[a]	XI.A.2.b (limitation 1[a])
8[b]	XI.A.2.c (limitation 1[b])
8[c]	XI.A.2.d (limitation 1[c])
8[d]	XI.A.2.e (limitation 1[d])
8[e]	XI.A.2.f (limitation 1[e])

4. Independent Claim 15

a. 15[Pre]: A method for generating speech output via a speech-enabled application, the method comprising

213. I understand that “[a] method for generating speech output via a speech-enabled application, the method comprising” is the preamble to claim 15 of the ’486 patent. I have been asked to assume that the preamble is a claim limitation. Under that assumption, in my opinion, *Walker* teaches the preamble.

214. *Walker* discloses that a TTS engine 18 synthesizes text into speech and provides the speech to voice application 16. *See* Ex. 1005 (*Walker*), ¶¶[0024] (“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.”), [0034] (“TTS

engine 18 then converts the cleaned text into speech and provides the speech which is labeled ‘Synthesized Audio’ in FIG. 1 to voice application 16. Voice application 16 then forwards or streams the speech to telephone user 12 in order to satisfy the information request from the telephone user.”). *Walker* discloses that one type of text that is synthesized is “a weather report for a particular city.” *See id.*, ¶[0024].

215. For the reasons I discussed in §XI.A.2.a (limitation 1[pre]), it is my opinion that a person of ordinary skill in the art would have understood *Walker*’s voice application 16 to be a speech-enabled application.

216. Accordingly, in my opinion, a person of ordinary skill in the art would have understood *Walker* to teach generating speech output via a voice application (“A method for generating speech output via a speech-enabled application”).

b. 15[a]: generating, using at least one computer system executing the speech-enabled application, a plurality of text strings, each of the plurality of text strings corresponding to a portion of a desired speech output;

217. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.

218. *Walker* discloses that text cleaner 24 receives raw text and a context identifier from context detector 26, and “then replaces, modifies, or otherwise

changes the raw text in accordance with the text cleaning rules to produce ‘Cleaned Text[.]’” Ex. 1005 (*Walker*), ¶[0033], FIG. 1. Accordingly, in my opinion, a person of ordinary skill in the art would have understood text cleaner 24 to be generating a “cleaned text” string from a raw text string using information about its context. *See id.*, ¶[0009] (“a method and system for converting raw text into cleaned text by modifying the raw text in accordance with its contexts”).

219. For the reasons I discussed in §XI.A.2.b (limitation 1[a]), it is my opinion that a person of ordinary skill in the art would have understood the weather report in the *Walker-Matsumoto* combination to include two or more sentences or text strings, each describing the weather for a different location/city or day.

220. *Walker* further discloses that the cleaned text from the text cleaner 24 is directed to a TTS engine 18 in TTS engine farm 19 for conversion or synthesis into speech. *See* Ex. 1005 (*Walker*), ¶[0034] (“Text cleaner 24 then provides the cleaned text to a TTS resource manager 28 which directs the cleaned text to an appropriate TTS engine 18 in TTS engine farm 19 for conversion or synthesis into speech.”).

221. As I discussed in §XI.A.1 (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.

222. Thus, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* combination to teach a system that generates a text weather report comprising one or more sentences ("plurality of text strings") to be synthesized into speech output.

223. In my opinion, a person of ordinary skill in the art would have also 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].

224. For the reasons I discussed in §XI.A.2.e (limitation 1[d]), it is my opinion that a person of ordinary skill in the art would have understood that the *Walker-Matsumoto* combination would have been implemented using a computer. Thus, in my opinion, a person of ordinary skill in the art would have also understood that the computer would have also executed the voice application (“the speech-enabled application”).

225. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* combination to teach a computer system executing a voice application (“using at least one computer system executing the speech-enabled application”) in which a text cleaner generates a cleaned version of a raw text weather report comprising two or more sentences (generating ... a plurality of text strings) wherein the cleaned text is provided to a TTS engine for conversion or synthesis into speech (“each of the plurality of text strings corresponding to a portion of a desired speech output”).

- c. 15[b]: inputting the plurality of text strings to at least one software module configured to identify a first portion of a first text string of the plurality of text strings as differing from a corresponding first portion of a second text string of the plurality of text strings, and a second portion of the first text string as not differing from a corresponding second portion of the second text string;**

226. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.

227. For the reasons I discussed in §XI.A.2.c (limitation 1[b]), the *Walker-Matsumoto* combination teaches “identify[ing] a first portion of a first text string of the plurality of text strings as differing from a corresponding first portion of a second text string of the plurality of text strings, and a second portion of the first text string as not differing from a corresponding second portion of the second text string.”

228. As I discussed herein, the *Walker-Matsumoto* combination also teaches that this functionality is performed by providing (“inputting”) the text weather report comprising one or more sentences (“the plurality of text strings”) into a TTS engine.

229. In my opinion, a person of ordinary skill in the art would have understood the plain and ordinary meaning of module in the

software/programming context to be “a collection of routines and data structures that performs a particular task or implements a particular abstract data type.” Ex. 1016 (*Computer Dictionary*), at 5.

230. Similarly, “modular software” was understood in the art to refer to “[a] program created from multiple standalone software components ... [that] can work together to perform the work for which the larger program is designed while still remaining individually usable—and reusable—in other programs.” *Id.*, 4.

231. In my opinion, a person of ordinary skill in the art would have understood the TTS engine of the *Walker-Matsumoto* combination to comprise at least one software module.

232. For example, *Matsumoto* illustrates its speech synthesizer as a functional block that receives text and outputs synthetic speech. *See* Ex. 1006 (*Matsumoto*), ¶[0037] (“The input unit 3 accepts text data of a sentence for generating a synthetic speech ... the waveform generation unit operate as a synthetic speech generation unit that generates the synthetic speech from the text data inputted to the input unit 3.”), FIG. 1. To be sure, while *Matsumoto* illustrates the speech synthesizer of figure 1 itself being comprised of additional functional blocks—which in turn can also comprise software modules—that would not alter a

person of ordinary skill in the art's understanding that the speech synthesizer as a whole is a software module.

233. This design—i.e., wherein the speech synthesizer is functionally self-contained and other components can call on its services without knowing how it works—is consistent with a person of ordinary skill in the art's understanding of a software module.

234. In my opinion, a person of ordinary skill in the art's understanding is further supported by *Walker* and the manner in which *Matsumoto*'s disclosure is incorporated therein.

235. For example, *Walker* shows the TTS engines 18 in its TTS engine farm 19 as distinct functional blocks that receive cleaned text and output synthesized audio output. *See* Ex. 1005 (*Walker*), FIG. 1. Thus, *Walker*, too, discloses its TTS engine(s)—which in the *Walker-Matsumoto* combination would be implemented using *Matsumoto*'s teachings, at least as it relates to synthesizing text weather reports—as autonomous functional block(s) that receive text input and output synthesized audio.

236. *Walker*'s and *Matsumoto*'s disclosures implementing a TTS engine as one or more software modules are also consistent with how person of ordinary skill

in the art typically implemented such functionality into a broader software/computer system.

237. For example, U.S. Pat. 7,779,357 (“*Naik*”) discloses an “audio user interface that generates audio prompts [to] help a user interact with a user interface of a computing device[.]” Ex. 1017 (*Naik*), Abstract. *Naik*’s system includes a “text-to-speech conversion engine 214” that “is a software module that converts text strings into audio files that can be played to generate a user interface audio prompt that audiblizes (verbalizes) a respective text string.” *Id.*, 5:51-55.

238. Additionally, U.S. Pat. No. 7,937,125 (“*May*”) discloses a wireless communication device that includes a “text-to-speech module 38.” Ex. 1018 (*May*) 5:60-64. *May* states that “the text-to-speech module 38 may be a software module that is run by the controller 25 when needed. Various text-to-speech conversion modules, which are known to those skilled in the art, may be used in accordance with the present invention.” *Id.*

239. Thus, in my opinion, a person of ordinary skill in the art would have understood that a TTS engine can be implemented as a software module in a broader software/computer system.

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

weather report comprising one or more sentences (“inputting the plurality of text stings”) into a TTS engine (“to at least one software module”).

- d. **15[c]: receiving, from the at least one software module, speech synthesis output to render the plurality of text strings with contrastive stress assigned to the first portion of the first text string and/or to the corresponding first portion of the second text string, but not to the second portion of the first text string, and not to the corresponding second portion of the second text string; and**

241. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.¹²

¹² I understand that to show that the feature “to render the plurality of text strings with contrastive stress assigned to the first portion of the first text string *and/or* to the corresponding first portion of the second text string” is taught by the prior art, it is sufficient to show that the prior art teaches *either* “to render the plurality of text strings with contrastive stress assigned to the first portion of the first text string” *or* “to render the plurality of text strings with contrastive stress assigned ... to the corresponding first portion of the second text string.”

242. For the reasons I discussed in §XI.A.2.e (limitation 1[d]), *Walker-Matsumoto*'s TTS engine generates “speech synthesis output to render the plurality of text strings with contrastive stress.”

243. As I discussed in §XI.A.4.c (limitation 15[b]), *Walker-Matsumoto*'s TTS engine is implemented as a software module.

244. As I discussed in §XI.A.2.d (limitation 1[c]), in the *Walker-Matsumoto* TTS engine, contrastive stress is “assigned to the first portion of the first text string ... but not to the second portion of the first text string, and not to the corresponding second portion of the second text string.”

245. *Walker* further discloses that the output of the TTS engine 18 is *received by* the voice application 16. *See* Ex. 1005 (*Walker*), ¶[0034] (“TTS engine 18 then converts the cleaned text into speech and provides the speech which is labeled ‘Synthesized Audio’ in FIG. 1 to voice application 16.”). Thus, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* combination to teach that output from a TTS engine is received by a voice application (“receiving, from the at least one software module”).

- e. **15[d]: generating, using the speech synthesis output, an audio speech output corresponding to the desired speech output.**

246. In my opinion, *Walker* in combination with *Matsumoto* teaches this feature.

247. For the reasons I discussed in §§XI.A.4.d (limitation 15[c]) and XI.A.2.e (limitation 1[d]), *Walker-Matsumoto*'s voice application receives from the TTS engine a synthesized audio of the weather report having contrastive stress.

248. Further, *Walker* discloses that the voice application provides the synthesized audio of the weather report to the telephone user 12:

TTS engine 18 [] ***provides*** the audio speech of the weather report text ***to telephone user 12 via voice application 16.***

Ex. 1005 (*Walker*), ¶[0024].

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. ***Voice application 16 then forwards or streams the speech to telephone user 12 in order to satisfy the information request from the telephone user.*** *Id.*, ¶[0034].

249. In my opinion, a person of ordinary skill in the art would have also understood the synthesized audio provided by the voice application to the telephone user to correspond to the desired speech output because *Walker* discloses

the telephone user using *Walker*'s system to request that a text weather report be outputted as synthesized audio.

250. Further, it is also my opinion that a person of ordinary skill in the art would have understood that “generating ... an audio speech output” as used in the context of element 15[d] encompasses *Walker-Matsumoto*'s voice application providing the synthesized audio of the weather report to the telephone user 12 by, for example, forwarding or streaming the audio to the user. *See id.*, ¶[0034] (“TTS engine 18 then converts the cleaned text into speech and provides the speech which is labeled ‘Synthesized Audio’ in FIG. 1 to voice application 16. ***Voice application 16 then forwards or streams*** the speech to telephone user 12 in order to satisfy the information request from the telephone user.”). This is similar to the ’486 patent’s disclosure which states that “[u]pon receiving speech output 290 from synthesis system 200, speech-enabled application 210 may play speech output 290 in audible fashion to user 212 as an output speech prompt. Speech-enabled application 210 may cause speech output 290 to be played to user 212 using any suitable technique(s), as aspects of the present invention are not limited in this respect.” Ex. 1001 (’486Pat), 19:20-26.

251. To be clear, it is my opinion that a person of ordinary skill in the art would have understood that element 15[d] does not encompass or refer to the

operation/actions of the synthesis system 200 in the '486 patent. That is because a person of ordinary skill in the art would have understood that “speech synthesis output” as used in the '486 patent refers to the synthesized audio output generated by the synthesis system:

At act 720, the generated text input may be input to one or more speech synthesis engines. At act 730, *speech output* corresponding to at least a portion of the text input *may be received from the speech synthesis engine(s). The speech output may include audio speech output including at least one portion carrying contrastive stress to contrast with at least one other portion of the audio speech output.* Method 700 ends at act 740, at which the audio speech output may be provided to one or more user(s) of the speech-enabled application.

Ex. 1001 ('486Pat), 39:10-18; *see also id.*, FIG.7.

252. Yet, the “generating” recited in element 15[d] “**uses** the speech synthesis output.” In other words, a person of ordinary skill in the art, in my opinion, would have understood that the “generating” in 15[d] does not refer to what is generated by the synthesis system 200 because it is **using** what is generated by the synthesis system (i.e., the speech synthesis output).

253. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto* combination to teach forwarding or streaming audio speech output representing a weather report to a user (“generating

... an audio speech output corresponding to the desired speech output”) using synthesized speech output from a TTS engine (“using the speech synthesis output”).

5. Dependent Claims 7 and 14

- a. **Claim 7: The method of claim 1, wherein the speech synthesis output comprises an indication of the first portion of the first text string and/or the corresponding first portion of the second text string as being assigned contrastive stress.**

254. In my opinion, *Walker* in combination with *Matsumoto* teaches this claim.¹³

255. In my opinion, a person of ordinary skill in the art would have understood that in the example weather report I described in §XI.A.2.c (limitation 1[b]), *Matsumoto*’s speech synthesizer increases the speed/duration, pitch, and

¹³ I understand that to show that this claim is taught by the prior art, it is sufficient to show that the prior art teaches *either* “the speech synthesis output comprises an indication of the first portion of the first text string ... as being assigned contrastive stress” *or* “the speech synthesis output comprises an indication of ... the corresponding first portion of the second text string as being assigned contrastive stress.”

volume using the assigned variation coefficients C1, C2, and C3 for the first portion “Kanto” (not the second portion “kyou no”) of the first text string (n=2) because “Kanto” is identified as different from “Tohoku” in the second text string (n=1).

256. Further, it is my opinion that a person of ordinary skill in the art would have understood these prosodic changes as being an indication of the assigned contrastive stress given by variation coefficients C1, C2, and C3.

257. Indeed, such prosodic changes in synthesized speech output is consistent with how the ’486 patent describes an indication of assigned contrastive stress to speech synthesis output:

Each portion of the text input labeled by the markers as carrying contrastive stress may be appropriately rendered as audio speech carrying contrastive stress, in accordance with the synthesis technique(s) used. ***The resulting speech output may exhibit increased parameters such as pitch, fundamental frequency, amplitude and/or duration during the portion(s) carrying contrastive stress, in relation to the baseline values of such parameters that would be exhibited by the same speech output if it were not carrying contrastive stress.***

Ex. 1001 (’486Pat), 37:57-66.

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

with increases to the speed/duration, pitch, and/or volume using variation coefficients C1, C2, and C3 for the identified first portion of the first text string—e.g., a(2, 2) (“Kanto”)—(“wherein the speech synthesis output comprises an indication of the first portion of the first text string ... as being assigned contrastive stress.”).

b. Claim 14: The at least one non-transitory computer-readable storage medium of claim 8, wherein the speech synthesis output comprises an indication of the first portion of the first text string and/or the corresponding first portion of the second text string as being assigned contrastive stress.

259. For the reasons I discussed in §XI.A.3.a (limitation 8[pre]), the *Walker-Matsumoto* combination teaches “at least one non-transitory computer-readable storage medium.”

260. For the reasons I discussed in §XI.A.5.a (Claim 7), the *Walker-Matsumoto* combination teaches “the speech synthesis output comprises an indication of the first portion of the first text string ... as being assigned contrastive stress.”¹⁴

¹⁴ I understand that to show that this claim is taught by the prior art, it is sufficient to show that the prior art teaches *either* “the speech synthesis output comprises an

B. Ground 2A: *Walker* in Combination With *Matsumoto* and *Malsheen* Renders Obvious Claims 2-5, 9-12, 16-19

- 1. Claim 2: The method of claim 1, wherein the identifying comprises identifying the first portion of the first text string as differing from the corresponding first portion of the second text string based at least in part on a normalized orthography of the first and second text strings.**

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

262. The '486 patent describes “normalized orthography” as encompassing the conversion of numerals into word form:

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*

indication of the first portion of the first text string ... as being assigned contrastive stress” *or* “the speech synthesis output comprises an indication of ... the corresponding first portion of the second text string as being assigned contrastive stress.”

text are known, and aspects of the present invention are not limited to any particular normalization technique.

Ex. 1001 ('486Pat), 15:6-18.

263. In my opinion, a person of ordinary skill in the art would have understood *Matsumoto* discloses that text weather reports can express temperature as a number represented by numerals. *See* Ex. 1006 (*Matsumoto*), ¶[0071] (“[]tomorrow’s lowest temperature in the Kansai region is 10 degrees.]”).

264. Thus, 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 at least two sentences in which the temperature is represented by numerals representing numbers, such as the example weather report illustrated below:

The temperature in the Kansai region is 70 degrees.

The temperature in the Tohoku region is 80 degrees.

265. To be sure, the above 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.

266. In my opinion, a person of ordinary skill in the art would have understood that weather reports include numbers such as a temperature for a particular location. *See* Ex. 1005 (*Walker*), ¶[0024] (“For example, telephone user 12 may request a weather report for a particular city.”); *see also* Ex. 1006 (*Matsumoto*), ¶[0071] (“[tomorrow’s [] lowest temperature in the Kansai region is 10 degrees.]”).

267. Furthermore, it is my opinion that a person of ordinary skill in the art would have understood that a weather report can include the weather for multiple locations.

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]).

Ex. 1006 (*Matsumoto*), ¶[0071].

268. Thus, in my opinion, a person of ordinary skill in the art would have understood that a weather report, like that shown in figure 6 of *Matsumoto*, can include numbers (e.g., temperature measurements, probability of precipitation, wind speeds, etc.) for different locations, rather than qualitative adjectives (e.g., “fair”). *See also* §XI.A.2.c (limitation 1[b]) (discussing that the weather report in

the *Walker-Matsumoto* combination can pertain to more than one location/city or day and comprise more than one sentence).

269. *Walker*, however, does not discuss in great detail how the context cleaner should modify or change numerals representing a number before the text is sent to the appropriate TTS engine.

270. *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.

271. *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., number, date, time, measurement units). *Id.*, 5:41-50; *see also* 6:65-7:3. 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).

272. For example, additional text processing may be performed by the “Number Expander procedure” which “expands each number value to a full word string based on the determination of the type of number represented by the text, and passes the result to the TTS converter” for conversion to speech. *Id.*, 7:59-62.

273. In my opinion, a person of ordinary skill in the art would have understood *Malsheen*'s reference to "expand[ing]" a number value to "a full word string" to mean that the number represented by numerals is written in words.

274. Further, it is my opinion that a person of ordinary skill in the art would have been motivated to incorporate *Malsheen*'s teaching that numerals representing a number (as opposed to, for example, a date or a zip code) should be converted to written words into the *Walker-Matsumoto* combination such that numbers appearing as numerals in the text weather report are converted to numbers written as words by the text cleaner 24 to ensure proper pronunciation by the speech synthesizer (i.e., "seventy" instead of "seven zero").

275. For instance, with respect to the above example weather report, the text cleaner 24 would modify the text to:

The temperature in the Kansai region is seventy degrees.

The temperature in the Tohoku region is eighty degrees.

276. Given the above example cleaned text input, in my opinion, a person of ordinary skill in the art would have understood that *Matsumoto*'s speech synthesizer would operate on the text strings as schematically illustrated below.

277. As shown, *Matsumoto*'s collation unit would identify, for example, that "eighty" is different from "seventy," each of which constitute a "portion" of

the first and second text strings, t(n=2) and t(n=1), respectively.” To be sure, a person of ordinary skill in the art would have understood that *Matsumoto*’s collation unit would have also identified that “Kansai” and “Tohoku” are different.

	b=1	b=2	b=3	b=4	b=5	b=6	b=7	b=8	b=9
t(n=1)	The	temperature	in	the	Kansai	region	is	seventy	degrees
t(n=2)	The	temperature	in	the	Tohoku	region	is	eighty	degrees

278. 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 numbers written as numerals into words before sending text to a TTS engine) to improve a similar system (*Walker-Matsumoto* teaches converting text weather reports incorporating numerals representing numbers to synthesized speech) in the same way (numerals representing numbers in the *Walker-Matsumoto* text weather reports are converted to words as taught by *Malsheen* before the weather report is sent to a TTS engine).

279. 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.

280. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach identifying that “eighty” in position $b=8$ in the newer sentence ($t(n=2)$) (“identifying the first portion of the first text string”) differs from “seventy” in position $b=8$ in the older sentence ($t(n=1)$) (“as differing from the corresponding first portion of the second text string”) based in part on converting numbers represented as numerals to numbers represented as words (e.g., “80” to “eighty” and “70” to “seventy”) (“based at least in part on a normalized orthography of the first and second text strings”).

2. Claim 3: The method of claim 1, wherein the first and second text strings represent different numerical fields within a larger text string.

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

282. As I discussed in §XI.A.2.b (limitation 1[a]), a person of ordinary skill in the art would have understood that a “text string” can be a word, phrase, or sentence.

283. As I discussed in §XI.B.1 (Claim 2), 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-Malsheen* combination's teaching:

The temperature in the Kansai region is 70 degrees.

The temperature in the Tohoku region is 80 degrees.

284. In my opinion, a person of ordinary skill in the art would have understood that in such an example weather report, the first text string (e.g., “80 degrees”) and second text string (e.g., “70 degrees”) are each numerical fields within a larger text string (e.g., n=2 and n=1).

285. Further, it is my opinion that a person of ordinary skill in the art would have understood text strings “80 degrees” and “70 degrees” to represent “numerical fields” because each provide a numerical value (i.e., “80” and “70”) that represent a particular field (i.e., temperature measurement of “degrees”).

286. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach that “80 degrees” and “70 degrees” (“wherein the first and second text strings”) are different numerical fields within a larger sentence (“represent different numerical fields within a larger text string”).

3. **Claim 4: The method of claim 3, wherein the numerical fields are selected from the group consisting of: currency fields, date fields, digit sequence fields, number fields, fractional number fields, ordinal number fields, telephone number fields, flight number fields, street number fields, time fields, and zipcode fields.**

287. For the same reasons I discussed in §XI.B.2 (Claim 3), *Walker-Matsumoto-Malsheen* combination teaches this claim, particularly that “the numerical fields are selected from the group consisting of: ... number fields.”

4. **Claim 5: The method of claim 1, wherein the receiving comprises receiving the first and second text strings as first and second parameters passed to a function called by the speech-enabled application to render the first and second text strings with a contrastive stress pattern.**

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

289. For the reasons I discussed in §XI.A.2.e (limitation 1[d]), it is my opinion that a person of ordinary skill in the art would have understood that the *Walker-Matsumoto* combination’s functionality would be implemented using a computer system.

290. As I discussed in §XI.A.3.a (limitation 8[pre]), it is my opinion that a person of ordinary skill in the art would have understood that this functionality would be performed by running software on the computer.

291. In my opinion, a person of ordinary skill in the art would have understood that *Malsheen*'s disclosure is also compatible with being implemented as software running on a computer. *See* Ex. 1007 (*Malsheen*), 1:19 (“The present invention is directed to computer applications”); 2:62-63 (“[T]he present invention is a computer system for converting a text message into an audio representation”); 3:65-4:5 (“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.”);

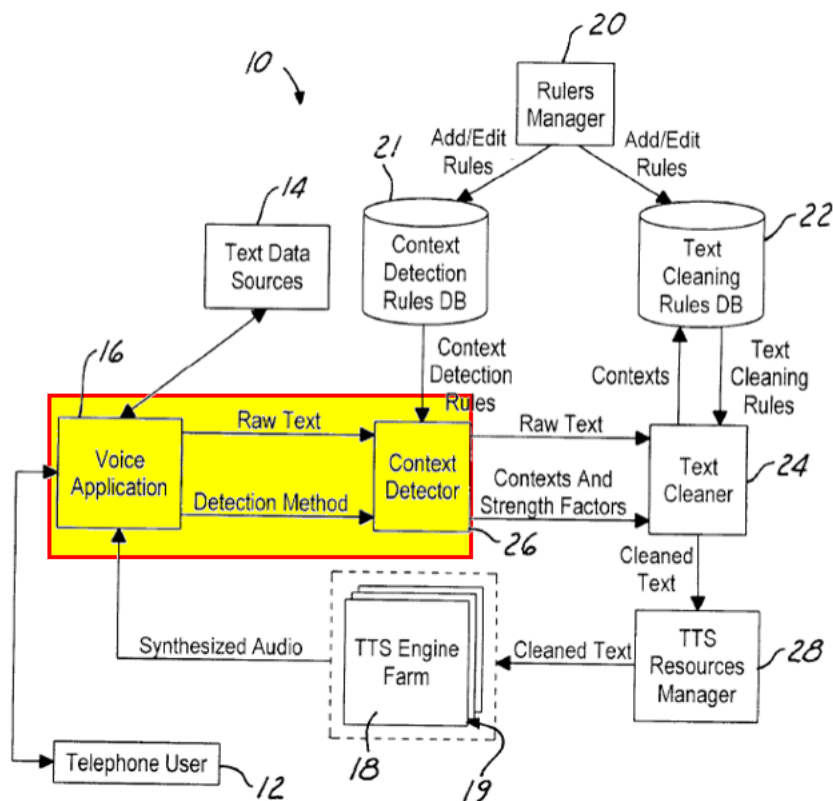
292. Further, it is my opinion that a person of ordinary skill in the art would have understood that the *Walker-Matsumoto-Malsheen* combination would have been implemented as various functional software blocks because each of *Walker*, *Matsumoto*, and *Malsheen* describe and organize their inventions by way of various functional blocks. *See* Ex. 1005 (*Walker*), FIG. 1; Ex. 1006 (*Matsumoto*), FIG. 1; Ex. 1007 (*Malsheen*), FIG. 1. *See also* §XI.A.4.c (limitation 15[b]).

293. For example, in my opinion, a person of ordinary skill in the art would have understood *Walker* to disclose each of context detector 26 and voice application 16 as different functional blocks within the wider software as a whole.

As I discussed in §XI.A.4.c (limitation 15[b]), *Walker* shows components of communication system 10 as distinct functional blocks which can comprise software modules.

294. Further, it is my opinion that a person of ordinary skill in the art would have understood a “function call” to be a request for the services of a particular function and that it includes the parameters needed for the function to perform its task. *See* Ex. 1016 (*Computer Dictionary*), at 3 (defining “function call” as “[a] program’s request for the services of a particular function. A function call is coded as the name of the function along with any parameters needed for the function to perform its task. The function itself can be a part of the program, be stored in another file and brought into the program when the program is compiled, or be a part of the operating system.”).

295. For the reasons I discussed in §XI.A.2.b (limitation 1[a]), in the *Walker-Matsumoto-Malsheen* combination, context detector 26 receives raw text from voice application 16. This is also shown in *Walker*’s figure 1 below.



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

296. Thus, in my opinion, a person of ordinary skill in the art would have understood the voice application 16 to be requesting that the context detector 26 perform a certain function (here, detect the context of the raw text data), and to provide to the context detector 16, *inter alia*, the raw text data as parameters for performing the requested function. See Ex. 1005 (*Walker*), ¶[0030] (“[T]he text document provided by voice application 16 to context detector 26 is labeled ‘Raw Text’ in FIG. 1 [because it] ... has not been processed.”), FIG 1. Accordingly, in

my opinion, *Walker*'s teaching is consistent with a person of ordinary skill in the art's understanding of one program function "calling" and "passing" data to another program function.

297. Accordingly, in my opinion, *Walker*'s teaching is consistent with a person of ordinary skill in the art's understanding of one program function "calling" and "passing" data to another program function.

298. Further, it is my opinion that a person of ordinary skill in the art would have further understood that in the example weather report discussed in §XI.B.1 (Claim 2), the first sentence would comprise a first parameter and the second would comprise a second parameter. I also note that *Matsumoto* discloses receiving multiple sentences one at a time. *See* Ex. 1006 (*Matsumoto*), ¶¶[0005], [0048]. Indeed, collation unit 14 stores sentences previously input in the accumulation unit 11 for comparative analysis. *See id.*, ¶[0048]. Thus, in my opinion, a person of ordinary skill in the art would have understood each sentence in the *Walker-Matsumoto* combination to be a parameter.

299. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach the context detector 26 receiving ("wherein the receiving comprises") the first and second sentences of the text weather report ("receiving the first and second text

strings”) as parameters passed by the voice application 16 to the context detector 26 (“as first and second parameters passed to a function called by the speech-enabled application”) as part of the process to apply contrastive stress to the synthesized audio of the weather report (“to render the first and second text strings with a contrastive stress pattern”).

5. Claims 9-12

300. For the reasons I discussed in §XI.A.3.a (limitation 8[pre]), the *Walker-Matsumoto* combination teaches “at least one non-transitory computer-readable storage medium.”

301. In my opinion, a person of ordinary skill in the art would have understood that *Malsheen*’s disclosure is also compatible with being implemented as computer-executable instructions stored on a non-transitory computer readable storage medium. See Ex. 1007 (*Malsheen*), 3:65-4:5 (“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.”), Claim 6 (“A **memory for storing data for access by programs being executed on a data processing system**, said memory comprising: ...”), Claim 9 (“A **memory for storing data for access**

by programs being executed on a data processing system, said memory comprising: ...”).

302. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Malsheen* combination to teach a non-transitory computer readable storage medium encoded with a plurality of computer-executable instructions for performing the functions recited in claims 9-12 for the same reasons as I described in §§XI.B.1 (Claim 2) - XI.B.4 (Claim 5), according to the table below.

Claim	Section
9	XI.B.1 (Claim 2)
10	XI.B.2 (Claim 3)
11	XI.B.3 (Claim 4)
12	XI.B.4 (Claim 5)

6. Claims 16-19

303. As I discussed in §XI.A.4.c (limitation 15[b]), a person of ordinary skill in the art would have understood that in the *Walker-Matsumoto* combination, the TTS engine as taught by *Matsumoto* would be implemented as a software module.

304. In my opinion, a person of ordinary skill in the art would have understood that nothing in the teachings of *Malsheen* that are incorporated into *Walker-Matsumoto* would alter this understanding.

a. Claims 16-18

305. 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 16-18 for the same reasons as I described in §§XI.B.1 (Claim 2) - XI.B.3 (Claim 4), according to the table below.

Claim	Section
16	XI.B.1 (Claim 2)
17	XI.B.2 (Claim 3)
18	XI.B.3 (Claim 4)

b. Claim 19

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

307. While claim 19's language differs slightly from that of claim 5 as shown in the table below, the scope and meaning of the two claims are not substantively different.

Claim 5	Claim 19
wherein the receiving comprises receiving the first and second text strings as first and second parameters passed to a function called by the speech-enabled application to render the first and second text strings with a contrastive stress pattern.	wherein the inputting comprises passing the first and second text strings as first and second parameters to a function called by the speech-enabled application to render the first and second text strings with a contrastive stress pattern.

308. As I discussed in §XI.B.4 (Claim 5), it is my opinion that a person of ordinary skill in the art would have understood that, in the *Walker-Matsumoto-Malsheen* combination, *Walker*’s context detector 26 would be “called” by the voice application 16, and that the voice application would transmit (“pass”) parameters, which the context detector would “receive,” comprised of the “Raw Text,” including the first and second text strings, that is being processed. Thus, the voice application passes (inputs) the first and second text strings as first and second parameters to the context detector.

309. Indeed, the ’486 patent itself makes clear that “passing” text to a function is the same as “inputting” text to a function:

For example, for the “221” portion of desired speech output 110, input code 120 could contain code that calls the name of a specific function for

mapping address numbers in English to sequences of audio filenames and passes the number “221” to that function *as input*.

Ex. 1001 (‘486Pat), 2:53-58.

Method 600 may be performed, for example, by a system executing a function to which the speech-enabled application *passes* fields of text representing portions of a desired speech output to be contrasted with each other, or by any other suitable system, machine and/or apparatus. Method 600 begins at act 610, at which input comprising a plurality of text strings may be *received* from a speech-enabled application.

Id., 38:25-31.

310. Accordingly, in my opinion, the *Walker-Matsumoto-Malsheen* combination teaches Claim 19 for the same reasons as I discussed in §XI.B.4 (Claim 5).

C. Ground 3A: *Walker* in Combination With *Matsumoto* and *Marple* Renders Obvious Claims 6, 13, and 20

- 1. Claim 6: The method of claim 1, wherein the speech synthesis output comprises identification of a plurality of audio recordings to render the plurality of text strings as speech, at least one of the plurality of audio recordings being selected to render the first portion of the first text string and/or the first portion of the second text string as speech carrying contrastive stress.**

311. In my opinion, *Walker* in combination with *Matsumoto* and *Marple* teaches this claim.¹⁵

312. For example, *Matsumoto* discloses generating a phonogram string from input text:

The linguistic processing unit 4 analyzes a dependency relation between the respective phrases that are detected, and outputs a result of this analysis as a phonogram string defined as a sentence that is segmented into a plurality of

¹⁵ I understand that to show that the feature “to render the first portion of the first text string *and/or* the first portion of the second text string as speech carrying contrastive stress” is taught by the prior art, it is sufficient to show that the prior art teaches *either* “to render the first portion of the first text string ... as speech carrying contrastive stress” *or* “to render ... the first portion of the second text string as speech carrying contrastive stress.”

words (phrase) and contains katakana characters representing the reading, accent information and symbols representing prosodeme.

Ex. 1006 (*Matsumoto*), ¶[0038].

313. The term “prosodeme” refers to a phoneme marked with its prosodic attributes, such as volume, pitch, and duration.

314. *Matsumoto* further discloses that the phonogram string is processed into a string of phonemes for which a normal phoneme length (speed/duration), pitch, and volume is calculated:

The pitch generation unit 6 *generates a [] phonemic string from the phonogram string* by a predetermined method.

Ex. 1006 (*Matsumoto*), ¶[0040].

The phoneme length generation unit 5 generates a phoneme length out of the phonogram string inputted from the linguistic processing unit 4 (step S3).

The phoneme length generation unit 5 *determines the phoneme length (normal phoneme length)* corresponding to the respective phonemes structuring the phonogram string.

Id., ¶[0097].

The pitch generation unit 6 has data for determining the pitch frequency corresponding to the phoneme, and *generates the pitch frequency (a normal pitch frequency)* on the basis of this data.

Id., ¶[0104].

The volume generation unit 7 *determines the volume (a normal volume)* for each phoneme of the new sentence from the pitch pattern and from the phoneme string.

Id., ¶[0105].

315. Additionally, as I discussed in §XI.A.2.d (limitation 1[c]), *Matsumoto* teaches later modifying the normal speed/duration, pitch, and volume for the phoneme string based on the calculated variation coefficients (C1, C2, C3) which are applied to the phonemes to generate synthetic speech with the assigned changes to prosody. Further, *Matsumoto* teaches that synthetic speech is generated by the waveform generation unit 8:

The *waveform generation unit 8 generates the synthetic speech* from the phoneme string, the phoneme length, the pitch pattern and the volume information, which are inputted from the volume information generation unit 7 (step S11). The waveform generation unit 8 outputs the thus-generated synthetic speech to the voice output device (not shown) such as the speaker connected to the speech synthesizer 1.

Ex. 1006 (*Matsumoto*), ¶[0107]; *see also* §XI.A.2.e (limitation 1[d]).

316. In my opinion, a person of ordinary skill in the art would have understood that phonemes are the sounds which make up a particular language. For example, *Etezadi* describes phonemes as sounds:

[A]ll languages do not share the same basic *phonemes, which are sets of sounds used to form syllables and ultimately words.*

Ex. 1011 (*Etezadi*), 1:19-21.

Referring still to FIG. 3, the English language phrase “The Beatles,” the name of a famous British music group, is broken into phonemes comprising the phrase in the English language table 310. *For example, the phonemes “th,” “e,” “b,” “ea,” “t,” “l,” and “s” are generated in table 310 for the English-language phrase “The Beatles.”*

Id., 7:13-18; *see also id.*, FIG. 3.

317. Further, it is my opinion that a person of ordinary skill in the art would also have understood that a phonogram string is a string of phonemes represented as symbols, as in a computer representation or a written form.

The linguistic processing unit 4 analyzes a dependency relation between the respective phrases that are detected, and outputs a result of this analysis as a phonogram string defined as a sentence that is segmented into a plurality of words (phrase) and contains katakana characters representing the reading, accent information and symbols representing prosodeme.

Ex. 1006 (*Matsumoto*), ¶[0038].

318. Thus, in my opinion, a person of ordinary skill in the art would have understood that in the *Walker-Matsumoto* combination, symbolic representations of sounds (i.e., phonemes, with associated pitch, amplitude, and duration) are generated from the input text of a weather report.

319. *Matsumoto*, however, does not discuss the details of how phonemes are actually realized as an audio waveform (sound) for a given phonogram string. Thus, in my opinion, a person of ordinary skill in the art would have been motivated to find details for how phonemes are realized as an audio waveform to implement *Matsumoto*'s teachings. In so looking, a person of ordinary skill in the art would have readily found *Marple*.

320. *Marple* discloses a "speech synthesizer for synthesizing speech from text" by concatenating or assembling phonemes. Ex. 1008 (*Marple*), ¶[0013]; *see also id.*, ¶¶[0043],[0051]. In particular, *Marple* aims to improve "synthetic, or 'machine' speech to 'humanize' it to sound more appealing and natural to the human ear" (*id.*, ¶[0037]) by providing "naturally rhythmic or musical speech and which can readily generate synthetic speech with one or more prosodies" (*id.*, ¶[0012]) using a database of pre-recorded phonemes (*see id.*, ¶¶[0013], [0056]).

321. For example, *Marple*'s synthesizer includes a text parser 10 that "parse[s] text to be synthesized into text elements expressible as phonemes" (*id.*, ¶[0013]) and provides this text to synthesis unit 12 (*see id.*, ¶[0051]). Synthesis unit 12 identifies and matches phonemes from a phoneme database to respective text elements and outputs the signal to wave generator 14. *See id.*, ¶¶[0043] (discussing that the synthesis unit "assemble[s] or concatenate[s] phonemes

selected from the database, in accordance with the output from the text parser ...”),
[0051] (“Synthesis unit 12 matches or corresponds suitable phonemes from
phoneme database 24 with respective text elements[,] assembles the phonemes[,]
and outputs the signal to wave generator 14.”).

322. *Marple* discloses that phonemes used for synthesizing speech from
input text are pre-recorded speech with natural/normal prosody stored in a
database. *See id.*, ¶[0050] (“The phonemes in phoneme database 26 have
parameters for a basic prosody model such as reportorial prosody model 28.”).

Marple additionally states:

One embodiment of a phoneme database useful in the practice of the
invention comprises a *single-prosodic, encoded recording of each of a
number of acoustic units constituting phonemes*. The encoded recordings
may comprise a basic phoneme set having a basic prosody. The single
prosody employed for the recordings may be a “*neutral*” *prosody*, for
example reportorial, or other desired prosody, depending upon the speech
synthesizer application. The phoneme set may be assembled, or constituted,
to serve a specific purpose, for example to provide a full range of a spoken
language, of a language dialect, or of a language subset suitable to a specific
purpose, for example an audio book, paper, theatrical work or other
document, or customer support.

Id., ¶[0056].

In one illustrative example ... a comprehensive *default database of 300-800 phonemes of various pitches, durations, and amplitudes is created from the recordings of about 10,000 sentences spoken by trained Lessac speakers.*

Id., ¶[0070].

323. *Marple* also discloses using prosody coefficients to change the default/normal pitch, amplitude, and duration of concatenated phonemes:

The invention includes embodiments wherein *suitable coefficients to transform the default prosody values in the database to alternative prosody values* are determined by mathematical calculation. In such embodiments, the prosody coefficients can be stored in a fast run-time database. This method avoids having to store and manipulate computationally complex and storage-hungry wave data files representing the actual pronunciations, as may be necessary with known concatenated databases.

Id., ¶[0069].

Wavelet warping by methods such as described above can be helpful in allowing prosody coefficients to be derived for transforming baseline speech to a desired alternative prosody speech in manner whereby the desired transformation can be obtained by simple arithmetical manipulation. For example, *changes in pitch, amplitude, and duration can be accomplished by multiplying or dividing the prosody coefficients.*

Id., ¶[0122].

324. Further, *Marple* discloses that after receiving the speech signal from synthesis unit 12, wave generator 14 “generate[s] the speech signal as a wave signal” (*id.*, ¶[0015], FIG. 1) by “employ[ing] wavelet transforms” (*id.*, ¶[0051]) that change “pitch, amplitude, and duration” (*id.*, ¶[0122]). These wavelet transforms implement prosody coefficients to adjust the pitch, amplitude, and duration of synthesized speech. *See id.*, ¶[0122].

325. In my opinion, a person of ordinary skill in the art would have found it obvious to incorporate the teachings of *Marple* into the *Walker-Matsumoto* combination because *Marple* provides additional details on how phonemes are stored and concatenated for generating synthetic speech with desired prosodic changes.

326. For example, in my opinion, a person of ordinary skill in the art would have appreciated that *Marple*’s teachings are directed to speech synthesis devices like the *Walker-Matsumoto* system to “introduc[e] greater human-like musicality or rhythm into synthesized speech” (Ex. 1008 (*Marple*), ¶[0113]) and to improve “synthetic, or ‘machine’ speech to ‘humanize’ it to sound more appealing and natural to the human ear” (*id.*, ¶[0037]); *See also id.*, FIG. 1.

327. In addition, a person of ordinary skill in the art would have understood that *Marple*’s teachings would improve the ability of the *Walker-*

Matsumoto system to “generate high quality speech” that is “resource-efficient” (*id.*, ¶[0012]) by using a synthesis unit that implements a phoneme database such that “at least 50 percent of the generated speech signal will be assembled from phonemes” (*id.*, ¶[0062]) which “avoids having to store and manipulate computationally complex and storage-hungry wave data files” (*id.*, ¶[0069]); *see also id.*, ¶[0051] (“Synthesis unit 12 matches or corresponds suitable phonemes from phoneme database 24 with respective text elements[,] assembles the phonemes[,] and outputs the signal to wave generator 14.”).

328. Indeed, a person of ordinary skill in the art would have understood that *Marple*’s phoneme database, which implements more “human-like speech,” would further improve the intelligibility of speech synthesized by the *Walker-Matsumoto* system. *See* §VIII.B (overview of *Walker*) (describing that *Walker* aims to correctly synthesize text associated with a particular type of document, such as a weather report), §VIII.C (overview of *Matsumoto*) (describing that *Matsumoto* aims to increase intelligibility and listener attention by emphasizing new words).

329. Furthermore, a person of ordinary skill in the art would have understood it was well-known that pre-recorded speech like *Marple*’s audio recordings are deployed in connection with wave generators that synthesize speech

and implement prosodic changes to phonemes, such as *Matsumoto*'s waveform generation unit. See Ex. 1008 (*Marple*), ¶[0122] (describing that "wave generator 14" assembles phonemes to change the "pitch, amplitude, and duration" of synthesized speech "by multiplying or dividing the prosody coefficients.").

330. Additionally, as I discussed in §XI.A.2.d (limitation 1[c]), *Matsumoto*'s waveform generation unit uses variation coefficients for speed/duration, pitch, and volume to correct synthesized speech for a phoneme string. In my opinion, a person of ordinary skill in the art would have understood *Marple*'s "prosody coefficients" to be analogous to *Matsumoto*'s "variation coefficients" because both are used to alter the baseline/normal prosody of phonemes generated from input text.

331. Thus, in my opinion, a person of ordinary skill in the art would have found it obvious to incorporate *Marple*'s teachings regarding its synthesis unit 12 and associated phoneme database 24 into *Matsumoto*'s speech synthesizer, as conceptually illustrated below.

phonemes) in the same way (generating synthetic speech from phonemes obtained from input text).

333. The combination would additionally have been obvious to a person of ordinary skill in the art because it would have amounted to applying a known technique (concatenating *Marple*'s audio recordings from a phoneme database) to a known device (*Walker-Matsumoto*'s speech synthesizer) ready for improvement to yield predictable results (to improve the naturalness and quality of synthesized speech).

334. Concatenating phonemes that are pre-recorded and stored in a database is consistent with how the '486 patent describes identifying audio recordings. *See* Ex. 1001 ('486*Pat*), 3:34-38 ("The concatenative TTS engine typically has access to a database of small audio files, each audio file containing a single subword unit (e.g., a phoneme or a portion of a phoneme) excised from many hours of speech pre-recorded by a voice talent."), 4:24-26; 10:60-64. The '486 patent also describes that "conventional full concatenative TTS synthesis" (*id.*, 4:6-23) involves "identify[ing] the audio recordings by ... concatenating the audio recordings to the speech-enabled application" (*id.*, 38:45-53).

335. Using the example weather report of §XI.B.1 (Claim 2), in my opinion, a person of ordinary skill in the art would have understood *Matsumoto*'s

speech synthesizer increases the speed/duration, pitch, and volume using the assigned variation coefficients C1, C2, and C3 for the phonemes comprising the word “Kanto” (i.e., the first portion of the first text string) because “Kanto” is identified as different from “Tohoku.” See §XI.A.5.a (Claim 7).

336. Thus, in my opinion, a person of ordinary skill in the art would have recognized that in the *Walker-Matsumoto-Marple* combination, the phoneme string comprising the word “Kanto” is first concatenated from a database of phoneme recordings and then receives contrastive stress based on the calculated variation coefficients C1, C4, and C3. See §XI.A.2.d (limitation 1[c]).

337. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Marple* combination to teach selecting audio recordings from a database of phonemes (“wherein the speech synthesis output comprises identification of a plurality of audio recordings to render the plurality of text strings as speech,”) and increasing the speed/duration, pitch, and volume using variation coefficients C1, C2, and C3 for the identified first portion of the first text string—e.g., a(2, 2) (“Kanto”)—(“at least one of the plurality of audio recordings being selected to render the first portion of the first text string ... as speech carrying contrastive stress”).

2. **Claim 13: The at least one non-transitory computer-readable storage medium of claim 8, wherein the speech synthesis output comprises identification of a plurality of audio recordings to render the plurality of text strings as speech, at least one of the plurality of audio recordings being selected to render the first portion of the first text string and/or the first portion of the second text string as speech carrying contrastive stress.**

338. In my opinion, *Walker* in combination with *Matsumoto* and *Marple* teaches this claim.

339. For the reasons I discussed in §XI.A.3.a (limitation 8[pre]), the *Walker-Matsumoto* combination teaches “at least one non-transitory computer-readable storage medium.”

340. In my opinion, a person of ordinary skill in the art would have understood that *Marple*’s disclosure is also compatible with being implemented as computer-executable instructions stored on a non-transitory computer readable storage medium:

*The disclosed invention can be implemented using various general purpose or special purpose computer systems One exemplary such computer system includes ... **computer readable storage media** ... for storing text, data, phonemes, speech and software or software tools useful for practicing the invention.*

Ex. 1008 (*Marple*), ¶[0141].

341. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Marple* combination to teach a non-transitory computer readable storage medium encoded with a plurality of computer-executable instructions for performing the functions recited in claims 13 for the same reasons as I described in §XI.C.1 (Claim 6).

3. **Claim 20: The method of claim 15, wherein the speech synthesis output comprises identification of a plurality of audio recordings to render the plurality of text strings as speech, at least one of the plurality of audio recordings being selected to render the first portion of the first text string and/or the first portion of the second text string as speech carrying contrastive stress.**

342. In my opinion, *Walker* in combination with *Matsumoto* and *Marple* teaches this claim for the same reasons as I described in §XI.C.1 (Claim 6).

D. Ground 1B: *Walker* in Combination With *Matsumoto* and *Bellegarda* Renders Obvious Claims 1, 7-8, 14-15

1. 1[c], 8[c], 15[c]

343. In my opinion, the *Walker-Matsumoto* combination teaches elements 1[c], 8[c], and 15[c] for the reasons I discussed in §§XI.A.2.d (limitation 1[c]), XI.A.3.b (limitation 8[c]), and XI.A.4.d (limitation 15[c]), respectively.

344. To the extent that Patent Owner argues and the Board agrees that deemphasizing a word/phrase constitutes assigning contrastive stress—and,

therefore, *Walker-Matsumoto* does not meet the portions of elements 1[c], 8[c], and 15[c] that require not assigning contrastive stress to the identified second portion of the first text string—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*’s TTS engine such that words that are the same (e.g., “kyuo no” at a(2, 1) as compared to “kyuo no” at a(1, 1)) are not deemphasized.

345. 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.

346. Further, as I noted in §XI.A.2.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.

347. As I discussed in §XI.A.2.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.

348. 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.

349. Indeed, only adjusting the prosody of words determined to be different between two text strings 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 text strings) for another (*Matsumoto* teaches

deemphasizing corresponding words that are the same between two text strings) to obtain predictable results (in the *Walker-Matsumoto-Bellegarda* combination corresponding words that different between two text strings would continue to receive contrastive stress).

350. 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 text strings would have reduced the computations required by the *Walker-Matsumoto-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 text strings.

351. 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 text strings) to improve similar devices (to reduce the computational requirements of *Walker-Matsumoto's* TTS engine by) in the same way (by not adjusting the prosody of words that are the same between two text strings in the *Walker-Matsumoto-Bellegarda* combination).

352. Accordingly, in my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-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 first text string—e.g., a(2, 2) (“Kanto”)—(“assigning contrastive stress to the identified first portion of the first text string”) but not changing the speed/duration, pitch, and volume for the second portion of the first text string—e.g., a(2, 1) (“Kyou no”)—(“but not to the identified second portion of the first text string”).

2. All Other Features of Claims 1, 7, 8, 14, and 15

353. In my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Bellegarda* combination to teach all other features of claims 1, 7, 8, 14, and 15 for the reasons I previously discussed according to the table below.

Element/Claim	Section
1[pre]	XI.A.2.a (limitation 1[pre])
1[a]	XI.A.2.b (limitation 1[a])
1[b]	XI.A.2.c (limitation 1b])
1[d]	XI.A.2.e (limitation 1[d])
1[e]	XI.A.2.f (limitation 1[e])
8[pre]	XI.A.3.a (limitation 8[pre])
8[a]	XI.A.2.b (limitation 1[a])

Element/Claim	Section
8[b]	XI.A.2.c (limitation 1[b])
8[d]	XI.A.2.e (limitation 1[d])
8[e]	XI.A.2.f (limitation 1[e])
15[pre]	XI.A.4.a (limitation 15[pre])
15[a]	XI.A.4.b (limitation 15[a])
15[b]	XI.A.4.c (limitation 15[b])
15[d]	XI.A.4.e (limitation 15[d])
7	XI.A.5.a (Claim 7)
14	XI.A.5.b (Claim 14)

E. Ground 2B (*Walker-Matsumoto-Bellegarda-Malsheen*)

354. In my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Bellegarda-Malsheen* combination to teach all features of claims 2-5, 9-12, and 16-19 for the reasons I previously discussed according to the table below.

Claim	Section
2	XI.B.1 (Claim 2)
3	XI.B.2 (Claim 3)
4	XI.B.3 (Claim 4)
5	XI.B.4 (Claim 5)
9	XI.B.1 (Claim 2)
10	XI.B.2 (Claim 3)
11	XI.B.3 (Claim 4)
12	XI.B.4 (Claim 5)
16	XI.B.1 (Claim 2)
17	XI.B.2 (Claim 3)
18	XI.B.3 (Claim 4)
19	XI.B.6.b (Claim 19)

F. Ground 3B (*Walker-Matsumoto-Bellegarda-Marple*)

355. In my opinion, a person of ordinary skill in the art would have understood the *Walker-Matsumoto-Bellegarda-Marple* combination to teach all features of claims 6, 13, and 20 for the reasons previously discussed according to the table below.

Claim	Section
6	XI.C.1 (Claim 6)
13	XI.C.2 (Claim 13)
20	XI.C.3 (Claim 20)

XII. SECONDARY CONSIDERATIONS

356. I am not aware of any evidence in the '486 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 ('486FH). To the extent Patent Owner asserts the existence of any secondary considerations in its responses, I reserve the right to address any such evidence.

357. 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