

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

Petition for *Inter Partes* Review of

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

TABLE OF CONTENTS

PETITIONERS' EXHIBIT LIST	v
LISTING OF CHALLENGED CLAIMS.....	vii
I. INTRODUCTION	1
II. GROUNDS FOR STANDING (37 C.F.R. §42.104(A))	1
III. STATEMENT OF PRECISE RELIEF REQUESTED FOR EACH CLAIM CHALLENGED	1
IV. OVERVIEW OF THE '486 PATENT	3
V. OVERVIEW OF THE PRIMARY PRIOR ART REFERENCES.....	6
A. <i>Walker</i> (Ex-1005).....	6
B. <i>Matsumoto</i> (Ex-1006)	10
VI. LEVEL OF ORDINARY SKILL IN THE ART	13
VII. CLAIM CONSTRUCTION	14
VIII. DETAILED EXPLANATION OF GROUNDS.....	15
A. Ground 1A: <i>Walker</i> in Combination With <i>Matsumoto</i> Renders Obvious Claims 1, 7-8, and 14-15	15
1. Motivation to Combine <i>Walker</i> and <i>Matsumoto</i>	15
2. Independent Claim 1	18
a. 1[pre].....	18
b. 1[a]	20
c. 1[b]	22
d. 1[c]	27

e.	1[d].....	34
f.	1[e].....	35
3.	Independent Claim 8	36
a.	8[pre].....	36
b.	8[a]-8[e]	39
4.	Independent Claim 15	39
a.	15[Pre]	39
b.	15[a].....	40
c.	15[b].....	42
d.	15[c]	45
e.	15[d].....	46
5.	Dependent Claims 7 and 14	48
a.	Claim 7.....	48
b.	Claim 14.....	49
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	50
1.	Claim 2.....	50
2.	Claim 3	54
3.	Claim 4.....	55
4.	Claim 5	56
5.	Claim 9-12.....	59
6.	Claims 16-19	60

a.	Claims 16-18.....	60
b.	Claim 19.....	61
C.	Ground 3A: <i>Walker</i> in Combination With <i>Matsumoto</i> and <i>Marple</i> Renders Obvious Claims 6, 13, and 20	62
1.	Claim 6	62
2.	Claim 13	70
3.	Claim 20	71
D.	Ground 1B (<i>Walker-Matsumoto-Bellegarda</i>)	71
1.	1[c], 8[c] and 15[c].....	71
2.	All Other Features of Claims 1, 7, 8, 14, and 15	75
E.	Ground 2B (<i>Walker-Matsumoto-Bellegarda-Malsheen</i>)	76
F.	Ground 3B (<i>Walker-Matsumoto-Bellegarda-Marple</i>).....	77
IX.	SECONDARY CONSIDERATIONS	77
X.	THE DISCRETIONARY FACTORS FAVOR INSTITUTING TRIAL	78
A.	35 U.S.C. §314(a).....	78
1.	Stay.....	78
2.	Trial Date	79
3.	Diligence/Investment	80
4.	Overlap.....	81
5.	Parties.....	81
6.	Other considerations.	81
B.	35 U.S.C. §325(d).....	82

XI.	MANDATORY NOTICES UNDER 37 C.F.R. §42.8.....	82
A.	Real Parties-in-Interest	82
1.	Related Matters	82
B.	Lead and Backup Counsel.....	83
C.	Service Information	83
D.	Power of Attorney	84
XII.	FEES	84

PETITIONERS' EXHIBIT LIST

Exhibit No.	DESCRIPTION
1001	U.S. Patent 8,825,486 (“’486Pat”)
1002	Declaration of Christopher M. Schmandt (“ <i>Schmandt</i> ”)
1003	Curriculum Vitae of Christopher M. Schmandt
1004	File History of U.S. Patent 8,825,486 (“’486FH”)
1005	U.S. Patent App. No. 2001/0049602 (“ <i>Walker</i> ”)
1006	U.S. Patent App. No. 2007/0233492 (“ <i>Matsumoto</i> ”)
1007	U.S. Patent 5,634,084 (“ <i>Malsheen</i> ”)
1008	U.S. Patent App. No. 2008/0195391 (“ <i>Marple</i> ”)
1009	U.S. Patent 7,010,489 (“ <i>Lewis</i> ”)
1010	U.S. Patent 7,177,807 (“ <i>Schmid</i> ”)
1011	U.S. Patent 8,290,775 (“ <i>Etezadi</i> ”)
1012	U.S. Patent 6,829,577 (“ <i>Gleason</i> ”)
1013	File History of U.S. Patent 8,447,610 (“’610FH”)
1014	U.S. Patent 7,313,523 (“ <i>Bellegarda</i> ”)
1015	U.S. Provisional Application No. 60/205,000
1016	Excerpts from MICROSOFT COMPUTER DICTIONARY (5th ed. 2002) (“ <i>Computer Dictionary</i> ”)
1017	U.S. Patent 7,779,357 (“ <i>Naik</i> ”)
1018	U.S. Patent 7,937,125 (“ <i>May</i> ”)

Exhibit No.	DESCRIPTION
1019	U.S. Patent 8,447,610 (“’610Pat”)
1020	Docket Control Order from <i>Cerence Operating Co. v. Samsung Electronics Co., Ltd. et al.</i> , No. 2:24-cv-00181-JRG-RSP (E.D. Tex.) (Dkt. 52)
1021	Cover Pleading of Plaintiff Cerence Operating Company’s Disclosure of Asserted Claims and Infringement Contentions (July 18, 2024)
1022	Summary of all patents/applications in the ’486 patent’s family

LISTING OF CHALLENGED CLAIMS

Claim 1	
1[pre]	A method for use with a speech-enabled application, the method comprising:
1[a]	receiving, from the speech-enabled application, input comprising a plurality of text strings;
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;
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;
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
1[e]	providing the speech synthesis output for the speech-enabled application.
Claim 2	
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.
Claim 3	

3	The method of claim 1, wherein the first and second text strings represent different numerical fields within a larger text string.
Claim 4	
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.
Claim 5	
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.
Claim 6	
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.
Claim 7	
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.
Claim 8	
8 <pre>[pre]</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:
8[a]	receiving, from the speech-enabled application, input comprising a plurality of text strings;
8[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;
8[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;
8[d]	generating speech synthesis output to render the plurality of text strings as speech having the assigned contrastive stress; and
8[e]	providing the speech synthesis output for the speech-enabled application.
Claim 9	
9	The at least one non-transitory computer-readable storage medium of claim 8, 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.
Claim 10	
10	The at least one non-transitory computer-readable storage medium of claim 8, wherein the first and second text strings represent different numerical fields within a larger text string.
Claim 11	

11	The at least one non-transitory computer-readable storage medium of claim 10, 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.
Claim 12	
12	The at least one non-transitory computer-readable storage medium of claim 8, 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.
Claim 13	
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.
Claim 14	
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.
Claim 15	
15[pre]	A method for generating speech output via a speech-enabled application, the method comprising:

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;
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;
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
15[d]	generating, using the speech synthesis output, an audio speech output corresponding to the desired speech output.
Claim 16	
16	The method of claim 15, wherein the at least one software module is configured to identify 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.
Claim 17	
17	The method of claim 15, wherein the first and second text strings represent different numerical fields within a larger text string.
Claim 18	
18	The method of claim 17, 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.
Claim 19	
19	The method of claim 15, 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.
Claim 20	
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.

I. INTRODUCTION

Petitioners Samsung Electronics Co., Ltd. and Samsung Electronics America, Inc. (collectively, “Petitioners” or “Samsung”) requests *inter partes* review of claims 1-20 of U.S. Patent No. 8,825,486 (“the ’486 patent”).

II. GROUNDS FOR STANDING (37 C.F.R. §42.104(A))

Petitioners certify that the ’486 patent is available for *inter partes* review and that Petitioners are not barred or estopped from requesting *inter partes* review to challenge the claims on the grounds herein.

III. STATEMENT OF PRECISE RELIEF REQUESTED FOR EACH CLAIM CHALLENGED

For purposes of this proceeding only, Petitioners apply the earliest alleged priority date of the ’486 patent: February 12, 2010.¹ Pre-AIA 35 U.S.C. §102 applies.

¹ The claims of the ’486 patent are not entitled to claim priority to the February 12, 2010 filing date of U.S. Patent 8,949,128 (*see* Ex. 1022) because the specification of the ’128 patent lacks, *inter alia*, written description support for at least limitations 1[b], 1[c], 8[b], 8[c], 15[b], and 15[c]. At best, the claims of the ’486 patent are entitled to a priority date no earlier than August 9, 2010 (i.e., the filing date of U.S. Patent 8,447,610). The Board need not resolve this issue because all of the prior art discussed in the Grounds predate February 12, 2010 by more than one year.

Petitioners respectfully request review and cancellation under 35 U.S.C. §311 of claims 1-20 (“Challenged Claims”) in combination with:²

Reference	Date(s)	Prior Art At Least Under
US20010049602 (“Walker”) (Ex-1005)	12/06/2001 (published)	§102(b)
US20070233492 (“Matsumoto”) (Ex-1006)	10/04/2007 (published)	§102(b)
US5634084 (“Malsheen”) (Ex-1007)	05/27/1997 (issued)	§102(b)
US20080195391 (“Marple”) (Ex-1008)	08/14/2008 (published)	§102(b)
US7,313,523 (“Bellegarda”) (Ex-1014)	12/25/2007 (issued)	§102(b)

Ground	Claims	Basis
1A	1, 7-8, 14-15	§103: <i>Walker, Matsumoto</i>
1B	1, 7-8, 14-15	§103: <i>Walker, Matsumoto, Bellegarda</i>
2A	2-5, 9-12, 16-19	§103: <i>Walker, Matsumoto, Malsheen</i>

² Petitioners do not concede that any Challenged Claims satisfy other requirements for patentability that cannot be raised in IPR, including 35 U.S.C. §§101 and 112.

Ground	Claims	Basis
2B	2-5, 9-12, 16-19	§103: <i>Walker, Matsumoto, Bellegarda, Malsheen</i>
3A	6, 13, 20	§103: <i>Walker, Matsumoto, Marple</i>
3B	6, 13, 20	§103: <i>Walker, Matsumoto, Bellegarda, Marple</i>

IV. OVERVIEW OF THE '486 PATENT

The '486 patent relates to “[t]echniques for generating synthetic speech with contrastive stress.”’*486Pat*, Abstract. 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. *See id.*, 8:25-28, 8:37-53, 12:41-47, 12:41-47.

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

Next, the system generates “an audio speech output corresponding to at least a portion of the text input” that contains “at least one portion carrying contrastive stress to contrast with at least one other portion of the audio speech output.” *Id.*, 4:51-54.

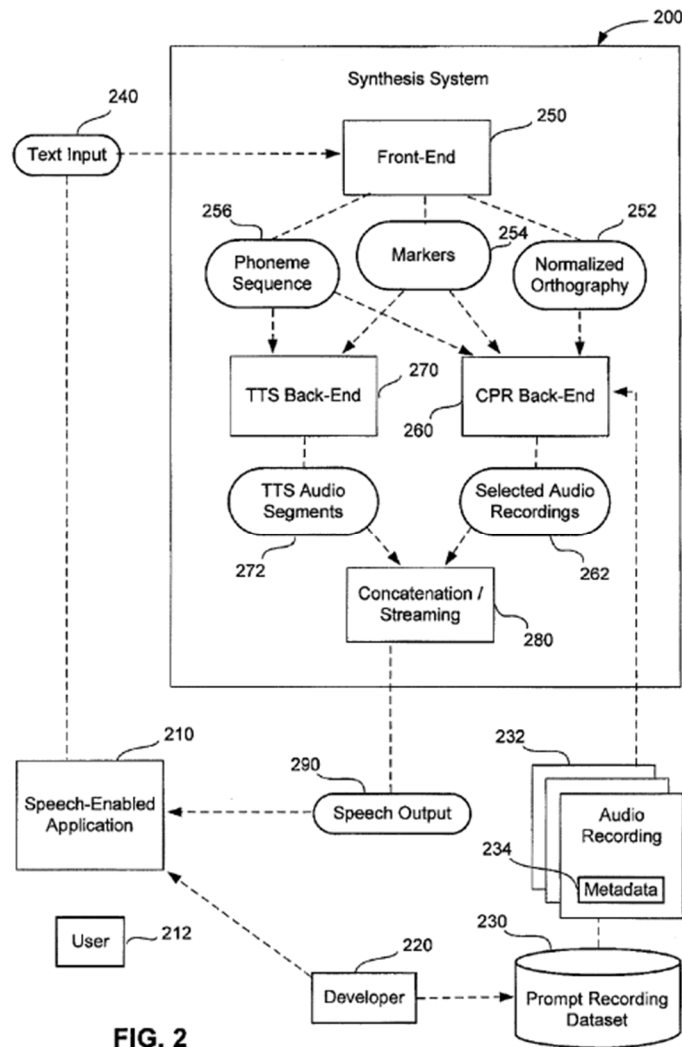
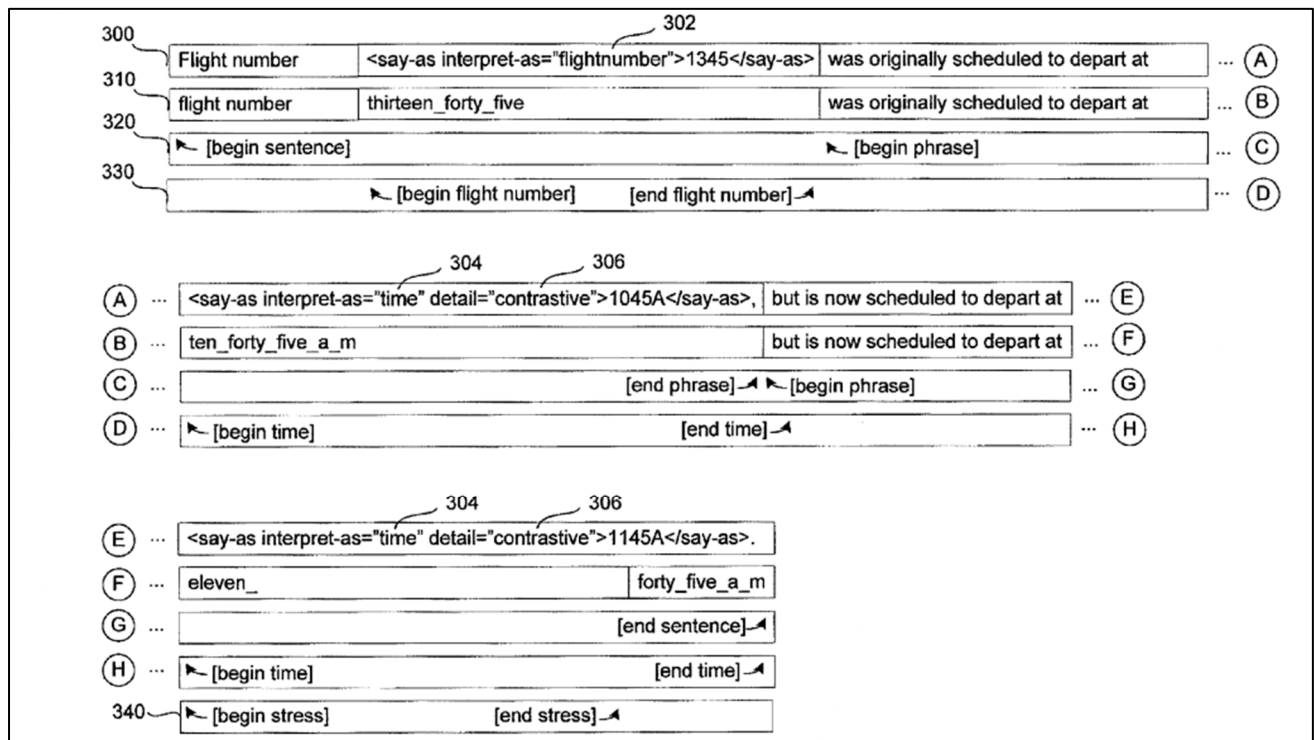


FIG. 2

'486Pat, FIG. 2.

Using the example of figure 3A below, “‘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).”’486Pat, 23:4-12.

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. *See* ’486Pat, 15:6-18. With reference to 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. Identifying the text normalization type aids synthesis system 200 to “identify[] portions of a text input that are meant to contrast with each other[.]” *Id.*, 24:25-29; *see also id.*, 16:25-27.



'486Pat, FIG. 3A.

V. OVERVIEW OF THE PRIMARY PRIOR ART REFERENCES

The prior art references analyzed in the grounds are analogous to the '486 patent because they are within the same field of endeavor as the '486 patent and are reasonably pertinent to one or more problems addressed by the '486 patent. *Schmandt* ¶¶43-48.

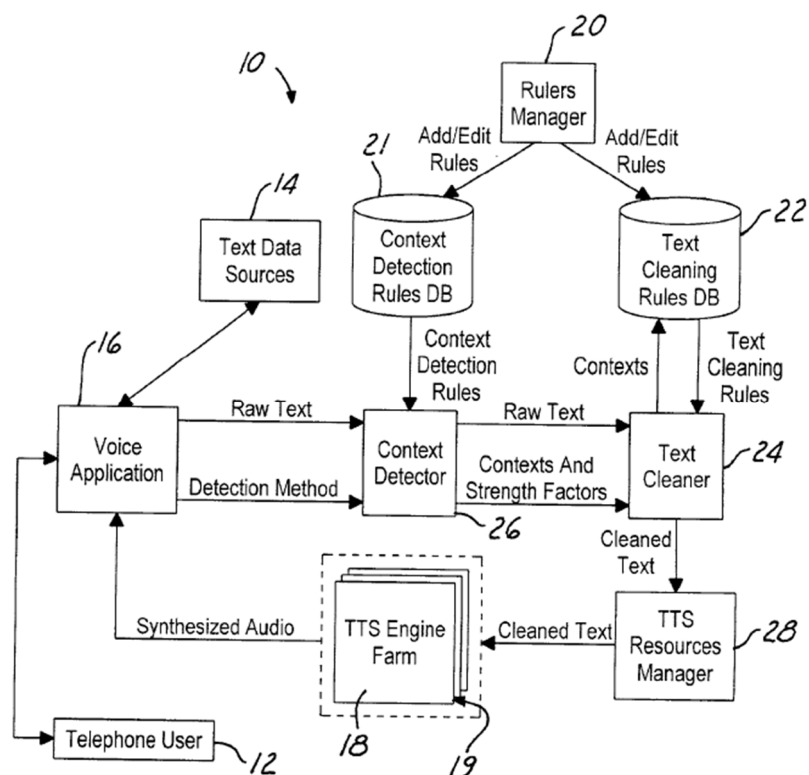
A. *Walker* (Ex-1005)

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.” *Walker* ¶[0002]. *Walker* explains that synthesized speech from TTS engines often results in

unnatural speech delivery due to the different contexts of words. *See id.*, ¶¶[0003]-[0005]. Thus, “without knowing the context of the document” the TTS engine may incorrectly convert text into speech. *Id.* 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]).

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. “Text data sources 14 may include ... *weather condition providers.*”³ *Id.*, ¶[0023].

³ All **bold**/*italics*/color emphases and annotations added unless noted otherwise.



Walker, FIG. 1.

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]. For example, a “telephone user 12 may request a weather report for a particular city” and, in response, “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.” *Id.*, ¶[0024].

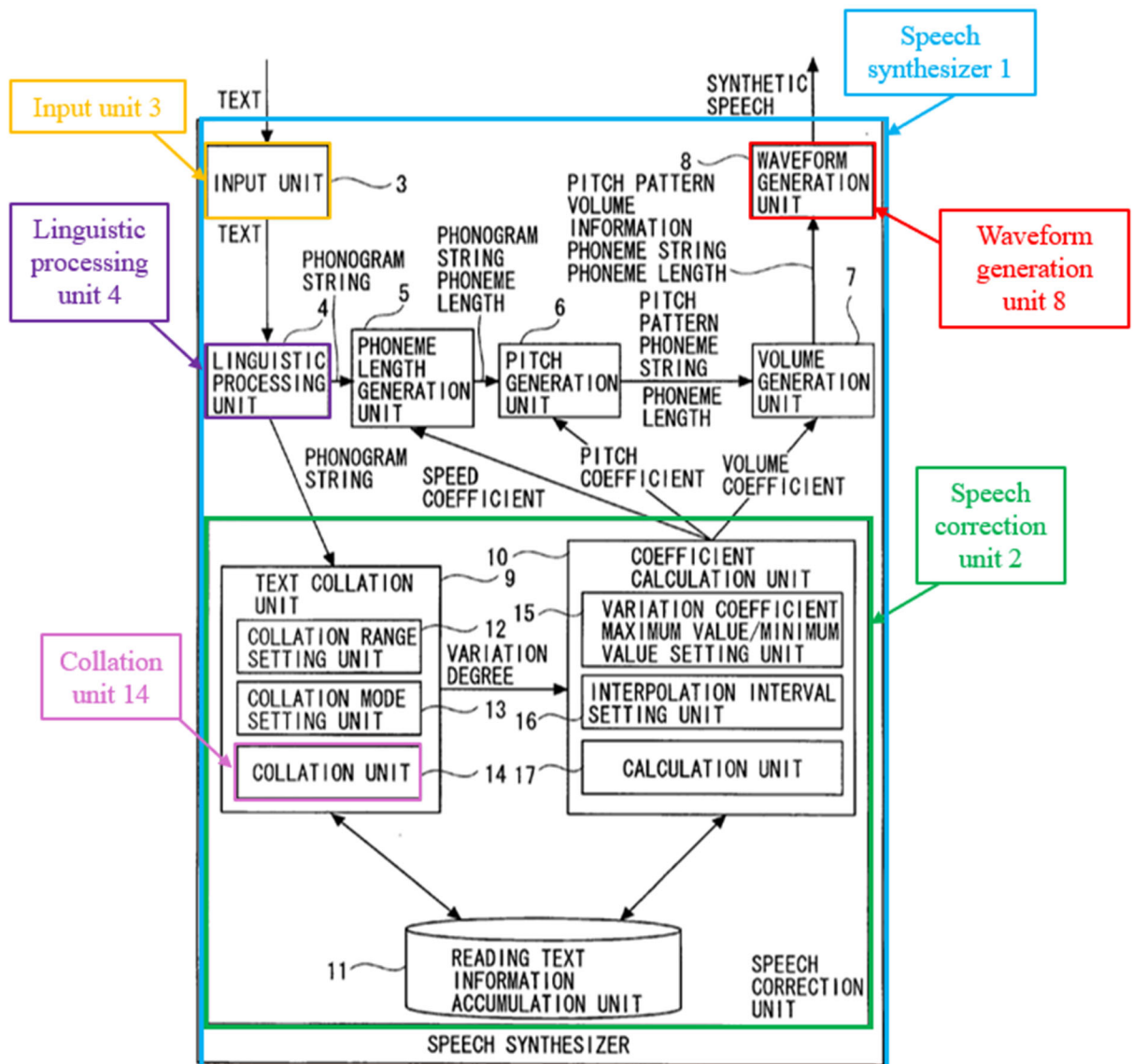
Walker further discloses preprocessing components “to determine the context of the text, i.e., context detection, and then modify[ing] the text based on the context of the text, i.e., text cleaning.” *Id.*, ¶[0025]. Specifically, *Walker*’s preprocessing includes a context detector 26 that uses “context 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 modifies “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]).

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]. For example, *Walker* discloses that “a TTS engine 18 ... 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].

B. *Matsumoto* (Ex-1006)

Matsumoto relates to “providing a hearer with [] easy-to-hear synthetic speech.” *Matsumoto*, Abstract. *Matsumoto* explains that in certain instances such as a “weather forecast,” speech synthesizers receive “almost the same sentences[.]” *Id.*, ¶[0005] (“[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.])”). However, if “the speech synthesizer vocalizes such sentences in a monotone” the hearer may fail to hear the “want-to-hear point.” *Id.* Accordingly, *Matsumoto*’s objective is to provide “the hearer with the easy-to-hear synthetic speech” (*id.*, ¶[0015]), that exhibits high naturalness by controlling prosodic changes in the sentence (*see id.*, ¶[0122]). Prosodic changes include variations in speed, pitch, and volume when words/phrases are synthesized into speech. *See id.*, ¶¶[0004], [0040]-[0042], [0103]-[0105]; *Schmandt* ¶61.

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



Matsumoto, FIG. 1 (annotated).

“[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]). 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].

Particularly, *Matsumoto* discloses that each sentence is assigned a sequential value “n” (n=1 being the oldest sentence) and each word or phrase is assigned a sequential value “b.” *See id.*, ¶¶[0072]-[0075]. 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

Matsumoto, FIG. 6.

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, “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.*

The calculated variation degree 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].

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].

VI. LEVEL OF ORDINARY SKILL IN THE ART

A person of ordinary skill in the art (“POSITA”) at the time of the alleged invention of the ’486 patent,⁴ would have possessed a Bachelor of Science (or

⁴ This analysis applies regardless of which of the two priority dates discussed in note 1 applies. *Schmandt* ¶23.

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. *Schmandt* ¶23.

VII. CLAIM CONSTRUCTION

Petitioners are unaware of any “prior claim construction determination” related to the ’486 patent. *See* 37 C.F.R. §42.100(b). Given the close correlation and substantial identity between the prior art references and the Challenged Claims, that no express constructions of the claims are necessary to assess whether the prior art reads on the Challenged Claims. Thus, the claims should be given their plain and ordinary meaning.⁵ *Phillips v. AWH Corp.*, 415 F.3d 1303 (Fed. Cir. 2005) (en banc).

⁵ Petitioners reserve all rights to raise claim construction arguments and other arguments in any parallel or future litigation concerning the ’486 patent. For example, comparing the claims to the accused products in the litigation may raise controversies that require construction of certain claim terms.

VIII. 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*

A POSITA would have been motivated to combine the teachings of *Walker* and *Matsumoto*, and would have had a reasonable expectation of success in doing so. *Schmandt* ¶¶78-94.

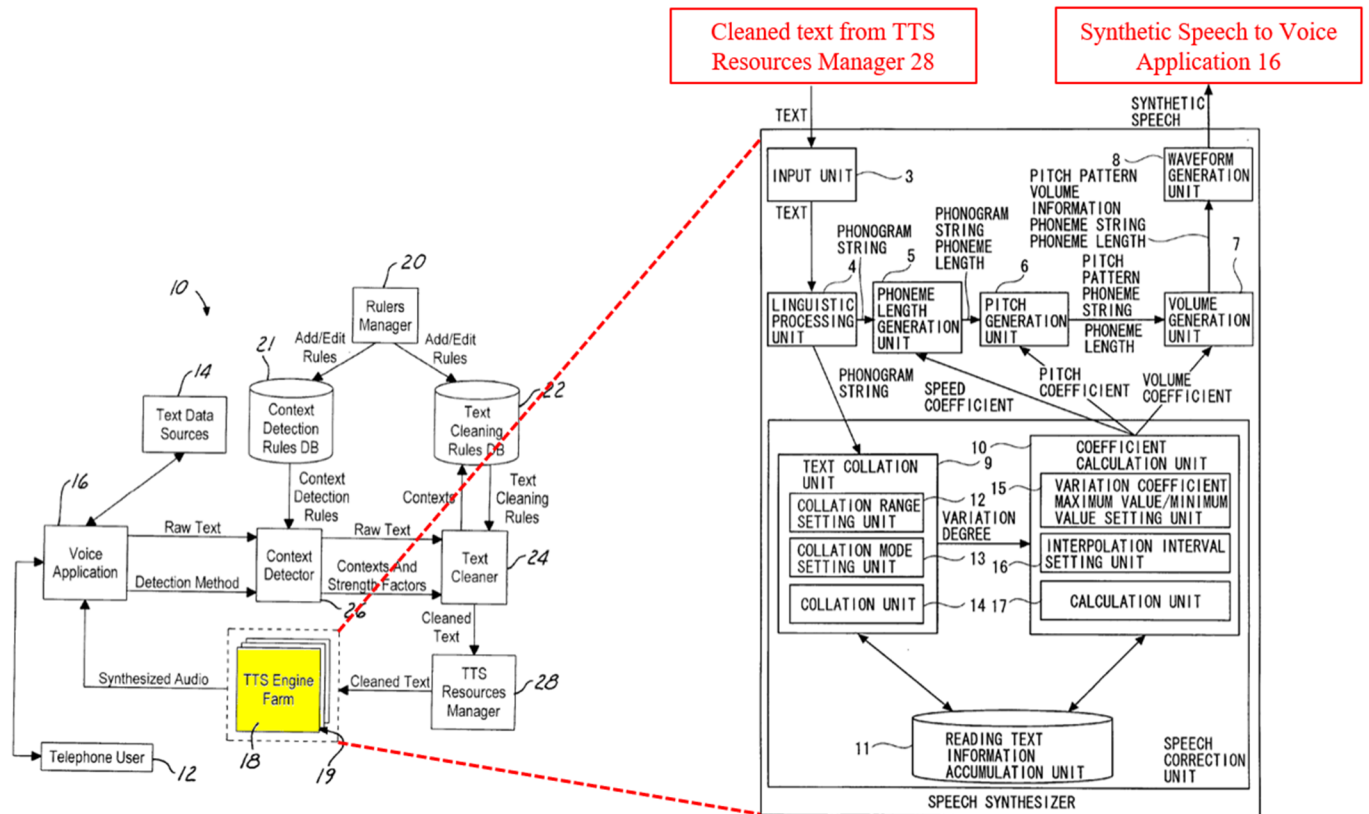
Walker teaches a system and method for improving the naturalness of synthesized speech using the text's context. *See* §V.A. *Walker* teaches obtaining more natural and higher quality speech synthesis by using a context detector and text cleaner to preprocess raw text received by a voice application. *See id.*, ¶¶[0030]-[0033]; *see also id.*, ¶¶[0011]-[0015], [0034]-[0038]. *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]). *Walker*, however, does not provide extensive implementation details regarding its TTS engines 18, 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. *Schmandt* ¶81; *See also Walker* ¶[0034]. Thus, a POSITA 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. *Schmandt* ¶82. In so looking, a POSITA would have readily found *Matsumoto*. *Schmandt* ¶82.

Matsumoto discloses a speech synthesizer that aims to provide easy-to-hear synthesized speech that exhibits high naturalness, for example, in a weather forecast. *See Matsumoto* ¶¶[0121]-[0122]; *see also id.*, ¶¶[0005], [0014], [0015], [0121]. It was well-known that speech synthesizers like *Matsumoto*'s were deployed in connection with speech applications, such as *Walker*'s voice application. *Schmandt* ¶84. 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. *See Walker* ¶[0024]; *Schmandt* ¶¶84-87 (discussing *Lewis and Schmid*).

A POSITA 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. *Schmandt* ¶¶88-90. Thus, a POSITA would have found it obvious to incorporate *Matsumoto*'s teachings regarding its speech synthesizer into *Walker* as one of *Walker*'s TTS engines 18 for converting text

weather reports into more naturally flowing, easy-to-hear synthesized speech, as conceptually illustrated in the figure below. *Schmandt* ¶91.



The combination would also have been obvious to a POSITA 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). *Schmandt* ¶92.

A POSITA 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 a POSITA's ability to implement in *Walker*. *Schmandt* ¶93. For example, both *Walker* and *Matsumoto* synthesize speech from documents. See *Walker* ¶¶[0004]-[0005], [0024], [0030]; *Matsumoto* ¶¶[0005], [0037]-[0038]. 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. *Schmandt* ¶94.

2. Independent Claim 1

a. 1[pre]

To the extent the preamble is limiting, *Walker* teaches this feature.⁶ *Schmandt* ¶¶95-101.

⁶ Petitioners use the term “teaches” as including both express teachings or those fairly suggested to a POSITA. *In re Baird*, 16 F.3d 380, 383

Walker discloses a method in which a telephone user 12 requests information and communication system 10 locates and converts the requested text into speech which is provided to the user via voice application 16. *See Walker* ¶¶[0018], [0023], FIG. 1. An example of a text to be converted to speech is a weather report. *See id.*, ¶[0023]. 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. *See id.*, ¶¶[0023]-[0024], [0034]. *Walker*'s voice application 16 is consistent with how the '486 patent describes a "speech enabled application." *See '486Pat*, 1:30-32 (describing a "speech-enabled application" as "capable of providing output to a human user in the form of speech."); *see also id.*, 1:35-57, 9:39-63; *Schmandt* ¶¶99-100.

Accordingly, *Walker* teaches a system having a voice application for performing a method of converting text into speech ("a method for use with a speech-enabled application"). *Schmandt* ¶101.

(Fed. Cir. 1994); *In re Keller*, 642 F.2d 413, 425 (CCPA, 1981) ("The test for obviousness is ... what the combined teachings of the references would have suggested to those of ordinary skill in the art." (citations omitted)).

b. 1[a]

Walker in combination with *Matsumoto* teaches this feature. *Schmandt* ¶¶102-115.

Walker discloses that context detector 26 receives raw text (i.e., unprocessed text) from voice application 16. *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 discussed in §VIII.A.2.a, a POSITA would have understood that voice application 16 is a speech-enabled application. Thus, a POSITA would have understood context detector 26 receives raw text from a speech-enabled application 16. *Schmandt* ¶103.

Walker 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, it would have been obvious to a POSITA that *Walker*’s weather report could pertain to more than one location/city or day. For example, *Matsumoto* discloses a weather forecast to include the weather for different locations and days.⁷ See *Matsumoto* ¶[0071] (describing

⁷ A POSITA would have understood weather reports and weather forecasts to include similar information. *Schmandt* ¶106.

“[t]oday’s weather” in the “Tohoku region” and “[t]omorrow’s weather” in the “Kansai region”); *see also id.*, ¶¶[0005], [0073]; *Schmandt* ¶¶104-105.

It would have been obvious to modify *Walker* based on *Matsumoto*’s disclosure such that the context detector 26 receives from voice application 16 raw text string(s) pertaining to the weather 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. *Schmandt* ¶107.

The combination would also have been obvious to a POSITA 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). *Schmandt* ¶108. Further, the combination would have been obvious to a POSITA 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). *Schmandt* ¶109.

Thus, a POSITA 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*. *Schmandt* ¶110. Indeed, a POSITA would have understood a "text string" to be a broad phrase that encompasses any combination of word(s), phrase(s), and/or sentence(s). *Schmandt* ¶¶111-113 (discussing *Etezadi* and *Gleason*). Indeed, this understanding is consistent with the '486 patent which describes an entire sentence as a "text string." *See* '486Pat, 19:36-40, FIG. 3A. *Schmandt* ¶114.

Accordingly, the *Walker-Matsumoto* combination teaches 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"). *Schmandt* ¶115.

c. 1[b]

Walker in combination with *Matsumoto* teaches this feature. *Schmandt* ¶¶116-145.

For the reasons discussed in §VIII.A.2.b, 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. *Schmandt* ¶117.

	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

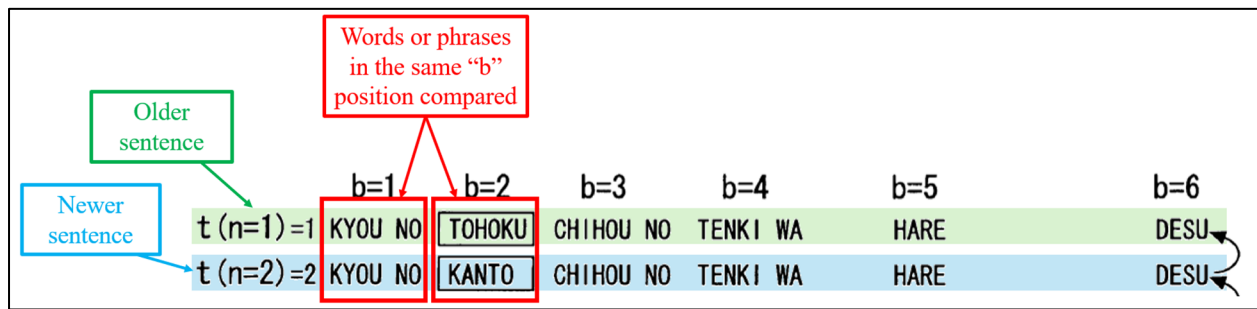
Matsumoto, FIG. 6.

While *Matsumoto*'s figure 6 shows five sentences (i.e., t(n=1) through t(n=5)), a POSITA would have understood *Matsumoto*'s system to be capable of operating on two or more sentences. *Schmandt* ¶118. For example, the variation degree of a given word is calculated by comparing it to the word in the same "b" position in the prior sentence. See *Matsumoto* ¶¶[0074], [0077]-[0079], Expression 1; *Schmandt* ¶¶119-120. *Matsumoto* further 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 *Matsumoto* ¶¶[0058]-[0059], [0061]-[0062], FIGs. 2-3; *Schmandt* ¶¶121-125.

The following discussion assumes 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.”

As discussed in §VIII.A.1, a POSITA would have found it obvious to incorporate *Matsumoto*’s TTS system into *Walker*’s TTS engine farm for synthesizing text weather reports into speech. *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 *Matsumoto* ¶¶[0044], [0074]; *Schmandt* ¶128. 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].

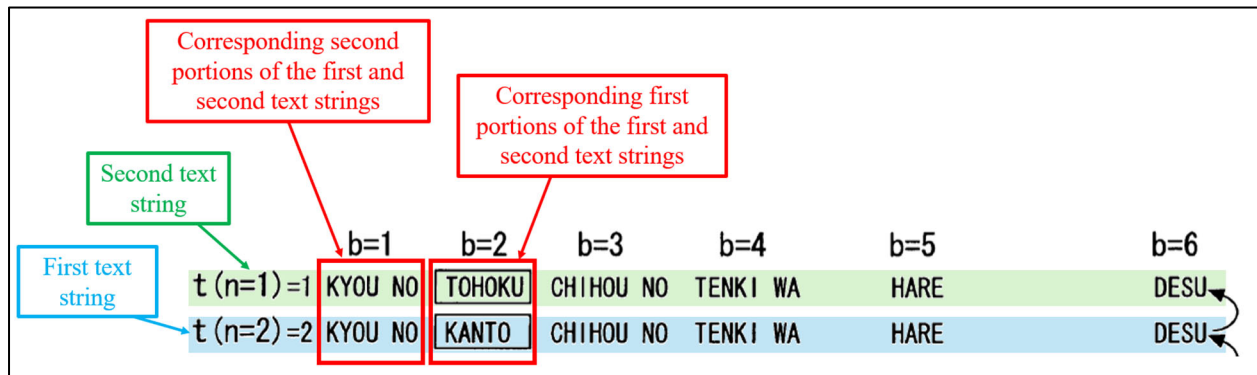


Matsumoto, FIG. 6 (excerpted and annotated).

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]. *Matsumoto* further discloses that when corresponding words or phrases 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; *Schmandt* ¶¶131-141 (discussing the calculation of the variation degree).

A POSITA 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. *Schmandt* ¶142. A POSITA 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 (i.e., $t(n=1)$ and $t(n=2)$). *Schmandt* ¶143.



Matsumoto, FIG. 6 (excerpted and annotated).

A POSITA would have further 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.” *Schmandt* ¶144.

Accordingly, the *Walker-Matsumoto* combination teaches 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”). *Schmandt* ¶145.

d. 1[c]

Walker in combination with *Matsumoto* teaches this feature. *Schmandt* ¶¶146-187.⁸

As discussed in §VIII.A.2.c, 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.”

Matsumoto discloses “calculation unit 17” uses the “variation degree obtained from the collation unit 14” to “calculate[] the variation coefficients (the speed coefficient, the pitch coefficient and the volume coefficient)” between corresponding words or phrases of two or more sentences. *Matsumoto* ¶[0054]. The variation coefficients C1 (speed), C2 (pitch), and C3 (volume), are used to adjust the

⁸ See *Ex parte Gross*, 2013 WL 6907805, *2 (PTAB Dec. 31, 2013) (noting that “‘and/or’ covers embodiments having element A alone, element B alone, or elements A and B taken together.”)

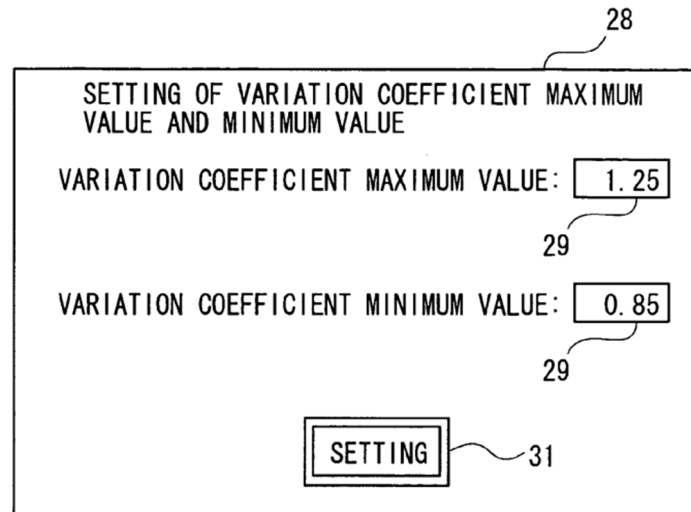
prosody of words/phrases in text when synthesized into speech. *See id.*, ¶¶[0103]-[0105].

A POSITA 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 Matsumoto* ¶¶[0010], [0011]-[0014], [0103]; *Schmandt* ¶¶150-155 (discussing *Bellegarda* and *Lewis*).

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 id.*, ¶[0054]. 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].

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). *See id.*, ¶¶[0064]-[0065]; *Schmandt* ¶158. *Matsumoto*’s figure 4 discloses an example in which the

variation coefficients (C1, C2, C3) have a user-set maximum value of 1.25 and minimum value of 0.85. *See Matsumoto* ¶[0091].



Matsumoto, FIG. 4.

When calculation unit 17 calculates a variation coefficient that is more than the user-set maximum value, the maximum variation coefficient value is used. *See id.*, ¶[0091]. 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. *See id.* A POSITA would have 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. *Schmandt* ¶¶163-165.

As discussed in §VIII.A.2.c, “Kanto” is identified as different from the corresponding word “Tohoku” and therefore, the calculated variation degree of “Kanto” is 1.0. Thus, a POSITA would have understood that *Matsumoto*’s system calculates variation coefficients (C1, C2, C3) for “Kanto” above 1.0. *Schmandt* ¶¶166-173 (discussing the calculation of the volume variation coefficient for “Kanto”). Accordingly “Kanto” (i.e., the first portion of the first text string) is emphasized. *Schmandt* ¶178.

As also discussed in §VIII.A.2.c, in the example weather report discussed in this ground, “kyuo no” in the second sentence (a(2, 1)) would be assigned a variation degree of “0,” because it is the same as “kyuo no” in the first sentence (a(1, 1)). When the variation degree is calculated to be “0,” the variation coefficient is also calculated to be “0.” *Schmandt* ¶¶174-177 (discussing the calculation of the volume variation coefficient for “kyou no”). Therefore, the user-set minimum variation coefficient is used, which in the case of figure 4, is 0.85. *Schmandt* ¶176. In this specific example in *Matsumoto*, words that are the same are deemphasized because the variation coefficients for speed/duration (C1), pitch (C2), and volume (C3) are less than 1.0. *Schmandt* ¶178. Additionally, a POSITA 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, a POSITA 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”). *Schmandt* ¶178.

Nevertheless, a POSITA would have understood that deemphasizing words/phrases that are the same (e.g., by decreasing the pitch, amplitude and/or duration), as described in this specific example in *Matsumoto*, 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. *Schmandt* ¶179. The ’486 patent describes “contrastive stress” as **increasing** the prosody for the words/phrases meant to contrast with each other. *See* ’486Pat, 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; *Schmandt* ¶180. 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. *Schmandt* ¶181.

Additionally, during prosecution of U.S. 8,447,610—which is the grand parent application of the '486 patent (*see* Ex. 1022)—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.” ’610FH, 83. 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.*” ’610FH, 147; ’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.” ’610FH, 147; ’486Pat, 8:25-28.

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.

'610FH, 83. Accordingly, both the specification and file history of the '486 patent are consistent with a POSITA's understanding that deemphasizing words that are the same is not an application of contrastive stress. *Schmandt* ¶185. Therefore, a POSITA would have understood that when *Matsumoto* deemphasizes words/phrases that are the same, *Matsumoto* does not assign "contrastive stress" because *Matsumoto* is not applying a stress to words/phrases *meant to contrast* with each other. *Schmandt* ¶186.

Accordingly, the *Walker-Matsumoto* combination teaches 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"). *Schmandt* ¶187.

e. 1[d]

Walker in combination with *Matsumoto* teaches this feature. *Schmandt* ¶¶188-195.

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). *See Matsumoto* ¶¶[0103]-[0105]. 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] (describing the waveform generation unit 8).

Each of *Walker* and *Matsumoto* disclose that their inventions are performed using a computer-based system. *Schmandt* ¶191. For example, *Walker* discloses that its “TTS engines ... synthesize the written words of a text document into **computer generated** speech understandable by humans.” *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**.”). Similarly, *Matsumoto* discloses that “[t]he present invention can be actualized as a program

that *makes a computer function as the speech synthesizer.*” *Matsumoto* ¶[0018]; see also *id.*, ¶¶[0036], [0093]. Thus, a POSITA 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. *Schmandt* ¶194.

Accordingly, the *Walker-Matsumoto* combination teaches 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”). *Schmandt* ¶195.

f. 1[e]

Walker in combination with *Matsumoto* teaches this feature. *Schmandt* ¶¶196-200.

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.” *Walker* ¶[0034]. As discussed in §VIII.A.1, a POSITA would have found it obvious to incorporate *Matsumoto*’s TTS system into *Walker*’s TTS engine farm for synthesizing text weather reports into speech. As discussed in §VIII.A.2.a,

a POSITA would have understood that voice application 16 is a speech-enabled application.

Accordingly, the *Walker-Matsumoto* combination teaches a TTS engine providing synthesized speech to a voice application (“providing the speech synthesis output for the speech-enabled application”). *Schmandt* ¶200.

3. Independent Claim 8

a. 8[pre]

To the extent the preamble is limiting, *Walker* alone or in combination with *Matsumoto*, teaches this feature. *Schmandt* ¶¶201-211.

For the reasons discussed in §VIII.A.2.a, *Walker* teaches “a method for use with a speech-enabled application.”

Walker discloses that “[t]ext-to speech (TTS) engines are **computing devices** which convert written text into audible **computer generated speech**.” *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**.”). Accordingly, a POSITA would have understood that *Walker*’s method is performed using software (i.e., computer readable instructions) running on a computer. *Schmandt* ¶204.

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, 5; *Schmandt* ¶205. A POSITA would have further understood that *Walker*’s computer comprises at least one processor for executing the software to implement *Walker*’s method. *Schmandt* ¶206 (describing *Bellegarda*). Additionally, a POSITA 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). *Schmandt* ¶207.

To the extent PO argues and the Board finds that *Walker* alone does not teach a non-transitory computer-readable storage medium for storing the software components/instructions executed by *Walker*’s computing device, this feature would have been obvious based on *Walker* in combination with *Matsumoto*. *Schmandt* ¶208.

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.***” *Matsumoto* ¶[0018]; see also *id.*, ¶¶[0036] (“[T]he 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], FIG. 10.

Thus, a POSITA 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. *Schmandt* ¶209. A POSITA 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). *Schmandt* ¶210.

Accordingly, the *Walker-Matsumoto* combination teaches 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”). *Schmandt* ¶211.

b. 8[a]-8[e]

Walker in combination with *Matsumoto* teaches Elements 8[a]-8[e] for the same reasons discussed in §§VIII.A.2.b-VIII.A.2.f, according to the table below. *Schmandt* ¶212.

Element	Section
8[a]	VIII.A.2.b
8[b]	VIII.A.2.c
8[c]	VIII.A.2.d
8[d]	VIII.A.2.e
8[e]	VIII.A.2.f

4. Independent Claim 15

a. 15[Pre]

To the extent the preamble is limiting, *Walker* teaches this feature. *Schmandt* ¶¶213-216.

For example, *Walker* discloses that a TTS engine 18 synthesizes text into speech and provides the speech to voice application 16. *See Walker* ¶¶[0024], [0034]. *Walker* discloses that one type of text that is synthesized is “a weather report for a particular city.” *Id.*, [0024]. For the reasons discussed in §VIII.A.2.a, a POSITA

would have understood *Walker*'s voice application 16 to be a speech-enabled application.

Accordingly, *Walker* teaches generating speech output via a voice application (“A method for generating speech output via a speech-enabled application”). *Schmandt* ¶216.

b. 15[a]

Walker in combination with *Matsumoto*, teaches this feature. *Schmandt* ¶¶217-225.

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[.]’” *Walker* ¶[0033]; *see also id.*, ¶[0009], FIG. 1. Accordingly, a POSITA would have understood text cleaner 24 to be generating a “cleaned text” string from a raw text string using information about its context. *Schmandt* ¶218.

For the reasons discussed in §VIII.A.2.b, a POSITA 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.

Walker further discloses that the cleaned text from the text cleaner 24 is directed to a TTS engine 18 for conversion or synthesis into speech. *See Walker* ¶[0034]. As discussed in §VIII.A.1, a POSITA would have found it obvious to incorporate *Matsumoto*'s TTS system into *Walker*'s TTS engine farm for synthesizing text weather reports into speech. Thus, a POSITA 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. *Schmandt* ¶222. A POSITA 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. *See Walker* ¶¶[0023]-[0024]; *Schmandt* ¶223.

For the reasons discussed in §VIII.A.2.e, a POSITA would have understood that the *Walker-Matsumoto* combination would have been implemented using a computer. Thus, a POSITA would have also understood that the computer would have also executed the voice application ("the speech-enabled application"). *Schmandt* ¶224.

Accordingly, the *Walker-Matsumoto* combination teaches 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”). *Schmandt* ¶225.

c. 15[b]

Walker in combination with *Matsumoto* teaches this feature. *Schmandt* ¶¶226-240.

For the reasons discussed in §VIII.A.2.c, 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.” As 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. *Schmandt* ¶228.

A POSITA 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.” *Schmandt* ¶229 (citing *Computer Dictionary*, 5). 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.” *Schmandt* ¶230 (citing *Computer Dictionary*, 4).

A POSITA would have understood the TTS engine of the *Walker-Matsumoto* combination to comprise at least one software module. *Schmandt* ¶231. For example, *Matsumoto* illustrates its speech synthesizer as a functional block that receives text and outputs synthetic speech.⁹ *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; *Schmandt* ¶232. This design—i.e., wherein the speech synthesizer is functionally self-contained and other

⁹ While *Matsumoto* illustrates the speech synthesizer of figure 1 itself being comprised of additional functional block—which in turn can also comprise software modules—that would not alter a POSITA’s understanding that the speech synthesizer as a whole is a software module. *Schmandt* ¶232.

components can call on its services without knowing how it works—is consistent with a POSITA’s understanding of a software module. *Schmandt* ¶233.

A POSITA’s understanding is further supported by *Walker* and the manner in which *Matsumoto*’s disclosure is incorporated therein. *Schmandt* ¶234. For example, *Walker* shows its TTS engines 18 as distinct functional blocks that receive cleaned text and output synthesized audio output. *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. *Schmandt* ¶235. *Walker*’s and *Matsumoto*’s disclosures implementing a TTS engine as one or more software modules are also consistent with how POSITA typically implemented such functionality into a broader software/computer system. *Schmandt* ¶¶236-238 (discussing *Naik* and *May*). Thus, a POSITA would have understood that a TTS engine can be implemented as a software module in a broader software/computer system. *Schmandt* ¶239.

Accordingly, the *Walker-Matsumoto* combination teaches 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”). *Schmandt* ¶240.

d. 15[c]

Walker in combination with *Matsumoto* teaches this feature. *Schmandt* ¶¶241-245.¹⁰

For the reasons discussed in §VIII.A.2.e, *Walker-Matsumoto*’s TTS engine generates “speech synthesis output to render the plurality of text strings with contrastive stress.” As discussed in §VIII.A.4.c, *Walker-Matsumoto*’s TTS engine is implemented as a software module. As discussed in §VIII.A.2.d, 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.”

Walker further discloses that the output of the TTS engine 18 is *received by* the voice application 16. See *Walker* ¶[0034]. Thus, a POSITA 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”). *Schmandt* ¶245.

¹⁰ The reasoning of note 8 applies.

e. 15[d]

Walker in combination with *Matsumoto* teaches this feature. *Schmandt* ¶¶246-253.

For the reasons discussed in §§VIII.A.4.d and VIII.A.2.e, *Walker-Matsumoto*'s voice application receives from the TTS engine a synthesized audio of the weather report having contrastive stress. Further, *Walker* discloses that the voice application provides the synthesized audio of the weather report to the telephone user 12. *See Walker* ¶¶[0024], [0034]. A POSITA 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. *Schmandt* ¶249.

A POSITA 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. *Schmandt* ¶250; *see also Walker* ¶[0034]. 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.” ’486*Pat*, 19:20-26; *Schmandt* ¶250.

To be clear, a POSITA 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. *Schmandt* ¶251. That is because a POSITA would have understood that “speech synthesis output” as used in the ’486 patent refers to the synthesized audio output generated by the synthesis system. See ’486*Pat*, 39:10-18, FIG.7; *Schmandt* ¶251. Yet, the “generating” recited in element 15[d] “*uses* the speech synthesis output.” In other words, a POSITA 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). *Schmandt* ¶252.

Accordingly, the *Walker-Matsumoto* combination teaches 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”). *Schmandt* ¶253.

5. Dependent Claims 7 and 14

a. Claim 7

Walker in combination with *Matsumoto* teaches this claim. *Schmandt* ¶¶254-258.

A POSITA would have understood that in the example weather report described in §VIII.A.2.c, *Matsumoto*'s speech synthesizer increases the speed/duration, pitch, and 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). *Schmandt* ¶255. A POSITA would have understood these prosodic changes as being an indication of the assigned contrastive stress given by variation coefficients C1, C2, and C3. *Schmandt* ¶256.

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. *See '486Pat.*, 37:57-66 ("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.”).

Schmandt ¶257.

Accordingly, the *Walker-Matsumoto* combination teaches 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.”). *Schmandt ¶258.*

b. Claim 14

For the reasons discussed in §VIII.A.3.a, the *Walker-Matsumoto* combination teaches “at least one non-transitory computer-readable storage medium.” *Schmandt ¶259.* For the reasons discussed in §VIII.A.5.a, the *Walker-Matsumoto* combination teaches “the speech synthesis output comprises an indication of the identified first portion of the first text string ... as being assigned contrastive stress.” *Schmandt ¶260.*

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

1. Claim 2

Walker in combination with *Matsumoto* and *Malsheen* teaches this feature. *Schmandt* ¶¶261-280.

The '486 patent describes “normalized orthography” as encompassing the conversion of numerals into word form. *See* '486*Pat*, 15:6-18 (“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.”); *Schmandt* ¶262

A POSITA would have understood *Matsumoto* discloses that text weather reports can express temperature as a number represented by numerals. *Schmandt* ¶263; *see also Matsumoto* ¶[0071] (“[The tomorrow’s [] lowest temperature in the Kansai region is 10 degrees.]”). Thus, a POSITA 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.

Schmandt ¶264. While the above example weather report is not explicitly stated in *Matsumoto*, a POSITA would have understood *Walker*'s disclosure would encompass, and apply to, such a weather report. *Schmandt* ¶265. Further, a POSITA would have understood that weather reports include numbers such as a temperature for multiple locations. *See Matsumoto* ¶¶[0024], [0071]; *Schmandt* ¶¶266-267. Thus, a POSITA 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"). *Schmandt* ¶268. *See also* §VIII.A.2.c.

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. *Schmandt* ¶269.

Malsheen discloses a text-to-speech converter for improving the handling of text that includes numbers. *See Malsheen*, 1:5-10. For example, *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). 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. A POSITA 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. *Schmandt* ¶273.

A POSITA 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”). *Schmandt* ¶274. 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.

Schmandt ¶275. Given the above example cleaned text input, a POSITA would have understood that *Matsumoto*’s speech synthesizer would operate on the text strings as schematically illustrated below. *Schmandt* ¶276. As shown, *Matsumoto*’s collation unit would identify 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.¹¹ *Schmandt* ¶277.

	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

A POSITA 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

¹¹ A POSITA would have understood that *Matsumoto*’s collation unit would also identify that “Kansai” and “Tohoku” are different. *Schmandt* ¶277.

to words as taught by *Malsheen* before the weather report is sent to a TTS engine). *Schmandt* ¶278. 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 a POSITA's ability to implement. *Schmandt* ¶279.

Accordingly, the *Walker-Matsumoto-Malsheen* combination teaches 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”). *Schmandt* ¶280.

2. Claim 3

Walker in combination with *Matsumoto* and *Malsheen* teaches this feature. *Schmandt* ¶¶281-286. As discussed in §VIII.A.2.b, a POSITA would have understood that a “text string” can be a word, phrase, or sentence. As discussed in §VIII.B.1, a POSITA 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.

A POSITA 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 $n=1$). *Schmandt* ¶284. Further, a POSITA 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”). *Schmandt* ¶285.

Accordingly, the *Walker-Matsumoto-Malsheen* combination teaches 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”). *Schmandt* ¶286.

3. Claim 4

For the same reasons discussed in §VIII.B.2, *Walker-Matsumoto-Malsheen* combination teaches this claim, particularly that “the numerical fields are selected from the group consisting of: ... number fields.” *Schmandt* ¶287.

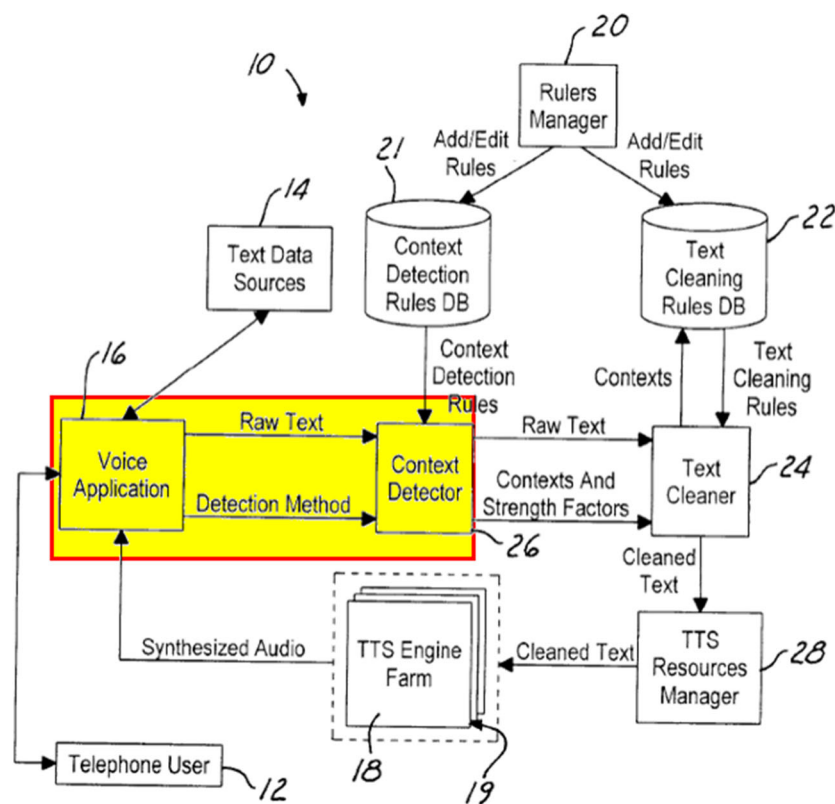
4. Claim 5

Walker in combination with *Matsumoto* and *Malsheen* teaches this feature. *Schmandt* ¶¶288-299.

For the reasons discussed in §VIII.A.2.e, a POSITA would have understood that the *Walker-Matsumoto* combination's functionality would be implemented using a computer system. And as discussed in §VIII.A.3.a, a POSITA would have understood that this functionality would be performed by running software on the computer. A POSITA would have understood that *Malsheen's* disclosure is also compatible with being implemented as software running on a computer. *See Malsheen*, 1:19 (“The present invention is directed to computer applications ...”); 2:62-63. 3:65-4:5; *Schmandt* ¶291.

A POSITA 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 Walker* FIG. 1; *Matsumoto* FIG. 1; *Malsheen* FIG. 1; *Schmandt* ¶292; *See also* §VIII.A.4.c. For example, a POSITA 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. *Schmandt* ¶293.

A POSITA 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. *Schmandt* ¶294 (citing *Computer Dictionary*, 3). For the reasons discussed in §VIII.A.2.b, in the *Walker-Matsumoto-Malsheen* combination, context detector 26 receives raw text from voice application 16, as shown below.



Walker, FIG. 1 (annotated)

Thus, a POSITA 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. *Schmandt* ¶296; see *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, *Walker*’s teaching is consistent with a POSITA’s understanding of one program function “calling” and “passing” data to another program function. *Schmandt* ¶297. A POSITA would have further understood that in the example weather report discussed in §VIII.B.1, the first sentence would comprise a first parameter and the second would comprise a second parameter. *Schmandt* ¶298; see also *Matsumoto* ¶[0005], [0048].

Accordingly, the *Walker-Matsumoto-Malsheen* combination teaches 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”). *Schmandt* ¶299.

5. Claim 9-12

For the reasons discussed in §VIII.A.3.a, the *Walker-Matsumoto* combination teaches “at least one non-transitory computer-readable storage medium.” A POSITA 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 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 ...”), claim 9 (same); *Schmandt* ¶301. Accordingly, the *Walker-Matsumoto-Malsheen* combination teaches 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 described in §§VIII.B.1-VIII.B.4, according to the table below. *Schmandt* ¶302.

Claim	Section
9	VIII.B.1
10	VIII.B.2
11	VIII.B.3

Claim	Section
12	VIII.B.4

6. Claims 16-19

As discussed in §VIII.A.4.c, a POSITA would have understood that in the *Walker-Matsumoto* combination, the TTS engine as taught by *Matsumoto* would be implemented as a software module. A POSITA would have understood that nothing in the teachings of *Malsheen* that are incorporated into *Walker-Matsumoto* would alter this understanding. *Schmandt* ¶304.

a. Claims 16-18

A POSITA 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 described in §§VIII.B.1-VIII.B.3, according to the table below. *Schmandt* ¶305.

Claim	Section
16	VIII.B.1
17	VIII.B.2
18	VIII.B.3

b. Claim 19

Walker in combination with *Matsumoto* and *Malsheen* teaches this claim. *Schmandt* ¶¶306-310.

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. *Schmandt* ¶307.

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.

As discussed in §VIII.B.4, a POSITA 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. *Schmandt* ¶308.

Indeed, the '486 patent itself makes clear that “passing” text to a function is the same as “inputting” text to a function. *See* '486*Pat*, 2:53-58 (“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*.”), 38:25-31 (“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 ... input comprising a plurality of text strings may be *received* from a speech-enabled application.”). *Schmandt* ¶309.

Accordingly, the *Walker-Matsumoto-Malsheen* combination teaches Claim 19 for the same reasons as discussed in §VIII.B.4. *Schmandt* ¶310.

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

1. Claim 6

Walker in combination with *Matsumoto* and *Marple* teaches this claim. *Schmandt* ¶¶311-337.

For example, *Matsumoto* discloses generating a phonogram string from input text. *See Matsumoto* ¶[0038]. *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. *See Matsumoto* ¶¶[0040], [0097], [0104]-[0105]. Additionally, as discussed in §VIII.A.2.d, *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.

Synthetic speech is generated by the waveform generation unit 8. *See Matsumoto* ¶[0107] (“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).”); *see also* §VIII.A.2.e.

A POSITA would have understood that phonemes are the sounds which make up a particular language. *Schmandt* ¶316 (discussing *Etezadi*). A POSITA 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. *See Matsumoto* ¶[0038]; *Schmandt* ¶317. Thus, a POSITA 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. *Schmandt* ¶318.

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

Marple discloses a “speech synthesizer for synthesizing speech from text” by concatenating or assembling phonemes. *See Marple* ¶¶[0013], [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 natural speech that can “generate synthetic speech with one or more prosodies” (*id.*, ¶[0012]) using a database of pre-recorded phonemes (*see id.*, ¶¶[0013], [0056]).

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] (synthesis unit

“assemble[s] or concatenate[s] phonemes selected from the database), ¶[0051]. (synthesis unit “matches or corresponds suitable phonemes from phoneme database 24 with respective text elements[,] assembles the phonemes[,] and outputs the signal to wave generator 14”). *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] (phonemes in the database “have parameters for a basic prosody model”), [0056], [0070]. *Marple* also discloses using prosody coefficients to change the default/normal pitch, amplitude, and duration of concatenated phonemes. *See id.*, ¶¶[0069] (“suitable coefficients [] transform the default prosody values in the database to alternative prosody values”), [0122].

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].

A POSITA would have found it obvious to incorporate *Marple*’s teachings 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. *Schmandt* ¶325. For example, a POSITA would have appreciated that *Marple*'s teachings are directed to speech synthesis devices like the *Walker-Matsumoto* system to introduce greater human-like rhythm into synthesized speech (see *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; *Schmandt* ¶326.

In addition, a POSITA 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]; *Schmandt* ¶327.

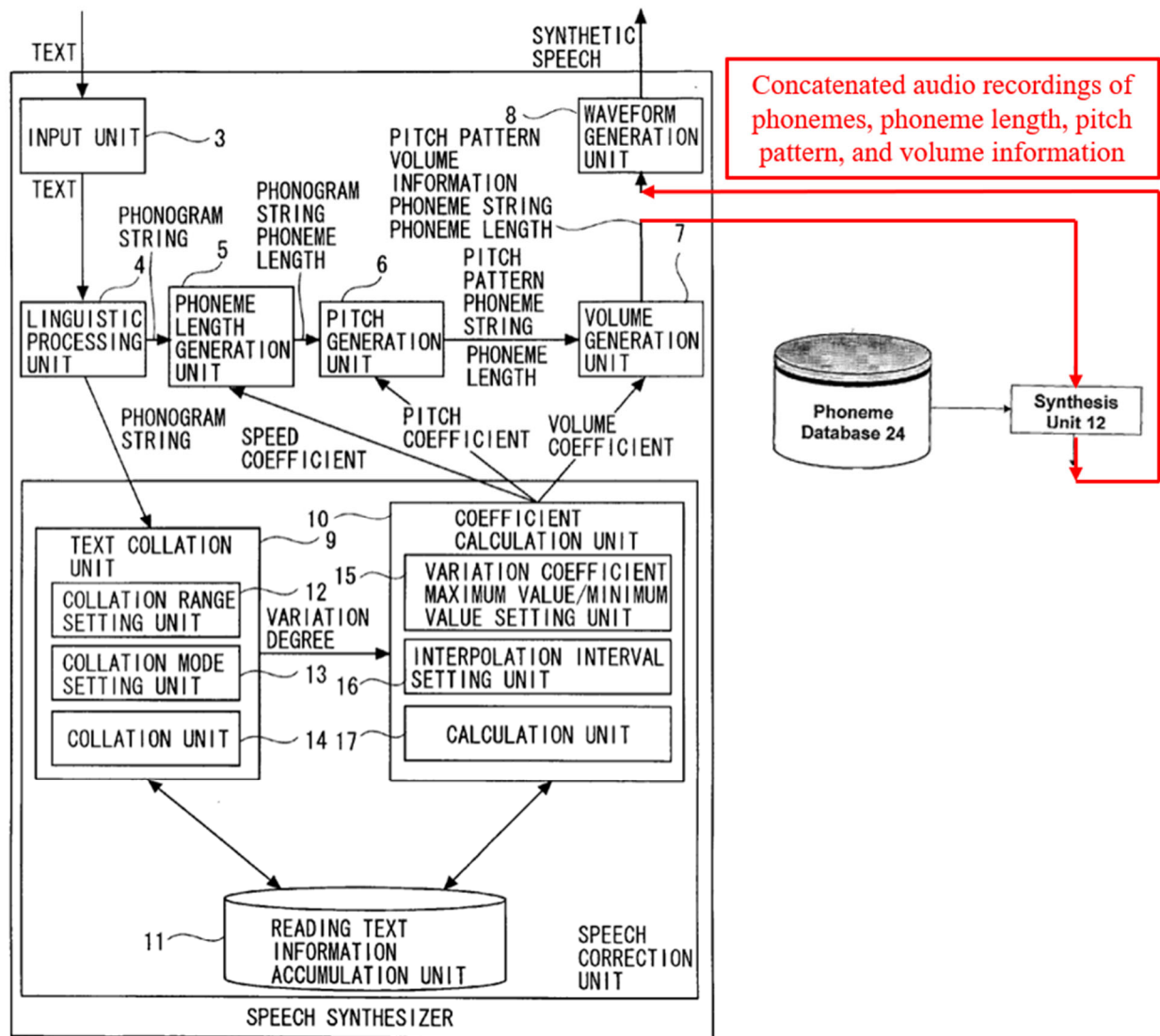
Indeed, a POSITA would have understood that *Marple*'s phoneme database, which implements more “human-like speech,” would further improve the intelligibility of speech synthesized by the *Walker-Matsumoto* system. *Schmandt* ¶328; see also §V.A (describing that *Walker* aims to correctly synthesize text associated with a particular type of document, such as a weather report), §V.B

(describing that *Matsumoto* aims to increase intelligibility and listener attention by emphasizing new words).

Furthermore, a POSITA 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. *Schmandt* ¶329; *see also 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.”).

Additionally, as discussed in §VIII.A.2.d, *Matsumoto*'s waveform generation unit uses variation coefficients for speed/duration, pitch, and volume to correct synthesized speech for a phoneme string. A POSITA 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. *Schmandt* ¶330.

Thus, a POSITA 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. *Schmandt* ¶331.



A POSITA would have also found the combination obvious because it amounted to the use of a known technique (*Malsheen* teaches a database of pre-recorded speech used to store phonemes for generating synthetic speech) to improve similar devices (*Walker-Matsumoto* teaches synthesizing speech from phonemes) in the same way (generating synthetic speech from phonemes obtained from input text).

Schmandt ¶332. The combination would additionally have been obvious to a POSITA because it 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). *Schmandt* ¶333.

Concatenating phonemes that are pre-recorded and stored in a database is consistent with how the ’486 patent describes identifying audio recordings. *Schmandt* ¶334. *See* ’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).

Using the example weather report of §VIII.B.1, a POSITA 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* §VIII.A.5.a.; *Schmandt* ¶335.

Thus, a POSITA 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* §VIII.A.2.d; *Schmandt* ¶336.

Accordingly, the *Walker-Matsumoto-Marple* combination teaches 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”). *Schmandt* ¶337.

2. Claim 13

Walker in combination with *Matsumoto* and *Marple* teaches this claim. *Schmandt* ¶¶338-341.

For the reasons discussed in §VIII.A.3.a, the *Walker-Matsumoto* combination teaches “at least one non-transitory computer-readable storage medium.” A POSITA 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. *See Marple* ¶[0141] (“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); *Schmandt* ¶340. Accordingly, the *Walker-Matsumoto-Marple* combination teaches 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 described in §VIII.C.1. *Schmandt* ¶341.

3. Claim 20

Walker in combination with *Matsumoto* and *Marple* teaches this claim for the same reasons as described in §VIII.C.1. *Schmandt* ¶342.

D. Ground 1B (*Walker-Matsumoto-Bellegarda*)

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

The *Walker-Matsumoto* combination teaches elements 1[c], 8[c], and 15[c] for the reasons discussed in §§VIII.A.2.d, VIII.A.3.b, and VIII.A.4.d, respectively. *Schmandt* ¶343.

To the extent that PO 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—these claim limitations would nevertheless have been obvious to a POSITA 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. *Schmandt* ¶344.

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. *See id.*, 5:60-67 (“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.”).

As noted in §VIII.A.2.d, *Matsumoto*’s maximum and minimum variation coefficient values are user-settable. Thus, it would have been obvious for a POSITA to set the minimum variation coefficient value to a value above 0, including 1.0. *Schmandt* ¶346. As further noted in §VIII.A.2.d, if the minimum variation coefficient is set to 1.0, there would be no change to the variation coefficients for words/phrases determined to be the same. Therefore, a POSITA would have understood that the minimum variation coefficient value could be set to a value of 1.0. *Schmandt* ¶347. A POSITA 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/phrases determined to be the same. *Schmandt* ¶348.

Only adjusting the prosody of words determined to be different between two text strings would have been obvious to a POSITA 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). *Schmandt* ¶349.

A POSITA 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. *Schmandt* ¶350. Thus, the combination would have also been obvious to a POSITA 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). *Schmandt* ¶351.

Accordingly, the *Walker-Matsumoto-Bellegarda* combination teaches 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”). *Schmandt* ¶352.

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

A POSITA would have understood the *Walker-Matsumoto-Bellegarda* combination to teach all other features of claims 1, 7, 8, 14, and 15 for the reasons previously discussed according to the table below. *Schmandt* ¶353.

Element/Claim	Section
1[pre]	VIII.A.2.a
1[a]	VIII.A.2.b
1[b]	VIII.A.2.c
1[d]	VIII.A.2.e
1[e]	VIII.A.2.f
8[pre]	VIII.A.3.a
8[a]	VIII.A.2.b
8[b]	VIII.A.2.c
8[d]	VIII.A.2.e
8[e]	VIII.A.2.f
15[pre]	VIII.A.4.a
15[a]	VIII.A.4.b

Element/Claim	Section
15[b]	VIII.A.4.c
15[d]	VIII.A.4.e
7	VIII.A.5.a
14	VIII.A.5.b

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

A POSITA 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 previously discussed according to the table below. *Schmandt* ¶354.

Claim	Section
2	VIII.B.1
3	VIII.B.2
4	VIII.B.3
5	VIII.B.4
9	VIII.B.1
10	VIII.B.2
11	VIII.B.3
12	VIII.B.4

Claim	Section
16	VIII.B.1
17	VIII.B.2
18	VIII.B.3
19	VIII.B.6.b

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

A POSITA 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. *Schmandt* ¶355.

Claim	Section
6	VIII.C.1
13	VIII.C.2
20	VIII.C.3

IX. SECONDARY CONSIDERATIONS

There is no 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* '486FH. To the extent PO

asserts the existence of any secondary considerations in its responses, Petitioners reserve the right to address any such evidence.

X. THE DISCRETIONARY FACTORS FAVOR INSTITUTING TRIAL

A. 35 U.S.C. §314(a)

To the extent the Patent Owner asks the Board to exercise its discretion to deny institution despite the strong invalidity showing on the merits, the Board should decline to do so because the weight of the factors articulated in *Apple Inc. v. Fintiv, Inc.*, IPR2020-00019, Paper 11 at 6 (PTAB Mar. 20, 2020) (precedential) favors institution.¹²

1. Stay

Factor 1 is neutral. Petitioners have not requested a stay but intend to do so. The PTAB has explained that it will not speculate on how any such motion would be resolved, before one is filed. *Google LLC v. Parus Holdings, Inc.*, IPR2020-00847, Paper 9 at 12 (PTAB Oct. 21, 2020).

¹² The *Fintiv* framework should not be followed because it is legally invalid as (1) exceeding the Director's authority, (2) arbitrary and capricious, and (3) adopted without notice-and-comment rulemaking.

2. Trial Date

The docket control order in the parallel litigation currently sets jury selection for January 5, 2026, about 12 months away. Ex. 1020. Any comparison of a projected FWD date against the scheduled trial date is speculative. *See Dish Network v. Broadband iTV*, IPR2020-01280, Paper 17 at 16 (PTAB Feb. 4, 2021) (“We cannot ignore the fact that the currently scheduled trial date is more than nine months away and much can change during this time”).

Setting that aside, this Petition is filed on January 21, 2025, so a FWD would be expected by July 2026. Although this is approximately six months after the currently scheduled (speculative) trial date, this factor is not determinative or considered in isolation. *Facebook, Inc. v. USC IP P’Ship*, IPR2021-00034, Paper 13 at 11 (PTAB April 13, 2021) (“[T]his factor is not considered in isolation, but holistically along with other factors”) (citation omitted). Moreover, the Board has instituted IPR and found, on similar facts, that this factor weighs only minimally in favor of denial. *See, e.g., Google LLC, et al. v. Multimodal Media LLC*, IPR2024-00056, Paper 9 at 8 (PTAB Apr. 12, 2024) (a period of about six months between trial and the expected FWD “weighs only marginally in favor” of denial); *NetNut Ltd. v. Bright Data Ltd.*, IPR2021-01492, Paper 12 at 9-16 (PTAB Mar. 21, 2022) (instituting IPR without stipulation and copending trial date six months before

FWD); *Facebook*, IPR2021-00034, Paper 13 at 11 (a period of five months between trial and expected FWD “slightly favors denial”); *Coolit Systems, Inc. v. Asetek Danmark A/S*, IPR2021-01195, Paper 10 at 11 (PTAB Dec. 28, 2021) (a period of five months between trial and expected FWD “weighs slightly in favor” of denial); *Equipmentsshare.com Inc. v. Ahern Rentals, Inc.*, IPR2021-00834, Paper 19 at (PTAB Nov. 16, 2021) (a period of seven months between trial and expected FWD “weighs somewhat in favor” of denial).

3. Diligence/Investment

Factor 3 weighs strongly against discretionary denial. No substantive orders have been issued by the court in the underlying litigation, and investment in the parallel litigation against Petitioner Samsung will have remained low at the time of institution. *See, e.g., Hulu LLC v. SITO Mobile R&D IP, LLC*, IPR2021-00298, Paper 11 at 12-14 (PTAB May 19, 2021) (holding that this factor supports instituting IPR given the early stage of the district court proceedings, the lack of substantial discovery related to invalidity claims, and the petitioner’s diligent filing of the petition after receiving preliminary infringement contentions). Assuming that a Decision on Institution is issued by July 2025, much work in district court will remain. *Id.* Claim construction briefing does not begin until May 23, 2025, and the

Markman hearing is scheduled for July 11, 2025. Opening expert reports are due August 11, 2025, and dispositive motions are due September 29, 2025. *Id.*

4. Overlap

Factor 4 weighs against discretionary denial. Petitioners challenge claims 1-20 in this Petition, whereas Patent Owner has alleged infringement of only claims 1-3, 5-7, 15-17, and 19-20. *See* Ex. 1021. and will likely reduce the number of asserted claims before trial. Accordingly, a material number of the challenged claims will not be addressed by the district court. *See, e.g., Precision Planting LLC v. Maschio Gaspardo S.p.A.*, IPR2024-00008, Paper 12 at 18 (PTAB Mar. 26, 2024) (holding that despite substantial overlap in issues between the IPR petition and parallel district court action, the inclusion of claims in the petition not contested in court argues against discretionary denial due to incomplete overlap).

5. Parties

Petitioners and PO are parties to the parallel litigation.

6. Other considerations.

The merits are compelling here. Petitioners therefore respectfully submit that the *Fintiv* factors favor institution and that discretionary denial of this Petition would be neither appropriate nor equitable.

B. 35 U.S.C. §325(d)

None of the references presented in the grounds herein were previously considered by the Office. Accordingly, the Board should not exercise its discretion under §325(d) to deny institution.

XI. MANDATORY NOTICES UNDER 37 C.F.R. §42.8

A. Real Parties-in-Interest

Petitioners identifies themselves as the real parties-in-interest.

1. Related Matters

To the best of Petitioners' knowledge, the '486 patent has only been involved in the following district court litigation: *Cerence Operating Co. v. Samsung Electronics Co., Ltd. et al.*, 2:24-cv-00181-JRG-RSP (E.D. Tex.), filed March 15, 2024.

To the best of Petitioners' knowledge, the '486 patent has not been challenged in any other *inter partes* review or post-grant review prior to this proceeding.

Petitioners are concurrently filing IPR2025-00458 directed to U.S. Patent No. 8,682,671, which is the parent of the '486 patent, and IPR2025-00460 directed to U.S. Patent No. 8,914,291, which is related to the '486 Patent. *See* Ex. 1022.

B. Lead and Backup Counsel

Lead Counsel	Back-Up Counsel
Ali R. Sharifahmadian, Reg. No. 48,202 (ali.sharifahmadian@arnoldporter.com) Arnold & Porter Kaye Scholer LLP 601 Massachusetts Ave., NW Washington, DC 20001-3743 Tel: 202-942-5000 Fax: 202-942-5999	Patrick Reidy (Reg. No. 72,148) patrick.reidy@arnoldporter.com Olivia Mathews (Reg. No. 80,801) olivia.mathews@arnoldporter.com Arnold & Porter Kaye Scholer LLP 70 West Madison Street Suite 4200 Chicago, IL 60602-4231 Tel: 312-583-2300 Fax: 312-583-2360 Douglas L. Clark (Reg. No. 68,443) douglas.clark@arnoldporter.com Arnold & Porter Kaye Scholer LLP 777 South Figueroa St., 44th Floor Los Angeles, CA 900017-5844 Tel: 213-243-4008 Fax: 213-243-4199

C. Service Information

Please address all correspondence to lead and back-up counsel at the addresses shown above. Petitioner consents to electronic service by e-mail at the following addresses:

ali.sharifahmadian@arnoldporter.com
patrick.reidy@arnoldporter.com

douglas.clark@arnoldporter.com
olivia.mathews@arnoldporter.com
xSamsungCerence2AP@arnoldporter.com

D. Power of Attorney

A power of attorney is filed herewith according to 37 C.F.R. §42.10(b).

XII. FEES

Petitioners concurrently electronically submits the required fees for this Petition. The Board is authorized to charge Arnold & Porter Kaye Scholer LLP's deposit account, No. 50-2387, for any fee deficiency.

Date: January 21, 2025

Respectfully submitted,

/Ali R. Sharifahmadian/
Ali R. Sharifahmadian (Reg. No. 48,202)
Counsel for Petitioner

CERTIFICATE OF COMPLIANCE

The undersigned hereby certifies that the foregoing Petition for *Inter Partes* Review contains 13,898 words, excluding those portions identified in 37 C.F.R. §42.24(a), as measured by the word-processing system used to prepare this paper.

/Ali R. Sharifahmadian/

Ali R. Sharifahmadian (Reg. No. 48,202)
Counsel for Petitioner

CERTIFICATE OF SERVICE

I certify that on January 21, 2025, I caused a true and correct copy of the foregoing Petition for *Inter Partes* Review of U.S. Patent No. 8,825,486 and supporting exhibits to be served via overnight delivery on the Patent Owner at the following correspondence address of record as listed on PAIR:

Douglas Wolf
Nuance Communications, Inc.
One Wayside Road
Burlington, MA 01803

A courtesy copy was also sent via electronic mail to Patent Owner's litigation counsel listed below:

Qi Tong: ptong@raklaw.com
James Tsuei jtsuei@raklaw.com
Benjamin Wang: bwang@raklaw.com
Andrew Weiss: aweiss@raklaw.com
Daniel Kolko: dkolko@raklaw.com
Minna Chan: mchan@raklaw.com
Shani Williams: swilliams@raklaw.com
Paul Kroeger: pkroeger@raklaw.com

/Ali R. Sharifahmadian/

Ali R. Sharifahmadian (Reg. No. 48,202)
Counsel for Petitioner