SBIR 23.2 Q&A Telecon Transcript
SOCOM232-002: Hokkien Low Density Language System

26 April 2023

SBIR Process Timeline
19 April 2023: Topics issued for pre-release
17 May 2023: USSOCOM begins accepting proposals via DSIP
31 May 2023: DSIP Topics Q&A closes to new questions at 12:00 PM ET
14 June 2023: Deadline for receipt of proposals no later than 12:00 PM ET

1. **Will Phase 1 cover cost of travel and expenses for data gathering in Tawain?**
   There are no travel costs included, or travel requirements included in any of SOCOM Phase Is.
2. **How is the problem addressed currently? What are the benefits of the proposed innovation when compared to the current solution?**
   Currently, we do not have a solution for Hokkien specifically. We are working on and have an initial set of languages for an edge-based translator that uses a laptop with GPU for processing and a phone for user interface. But there is no current solution that we have for an edge capability for Hokkien.
3. **Are there any specific form factor requirements or constraints for the portable device, such as size, weight, or battery life?**
   Right now, what we have for a form factor with our 3-3 radiant large density language, for lack of a better term, is a computer, a tensor book, as the portable cloud, and then we're using an S22 phone as the interface. We are looking to shrink that down to a smaller size. But as it stands right now, the main things that we're looking for guidance is that it can be portable.
4. **How is the problem addressed currently?**
   We're trying to develop a library of languages for the Special Operations community. We are currently tackling the large density languages like Chinese, Russian, Arabic, French, Spanish, Ukrainian, et cetera, but we also realize that a lot of our partner forces are going to probably be speaking lower density languages. We want to start to develop a model or a template for getting after low density languages, and we thought Hokkien would be a great place to start, since it is so significantly challenging to do as a traditionally unwritten language.
5. **Will there be opportunities for collaboration or partnership with other organizations, researchers, or subject matter experts during the project?**
   We would be happy to have any proposers work with the companies we currently have that are language SMEs that are working on our current set of languages that work on the end-to-end solution.

SBIR Program Office: Subcontracting wise, you have up to 1/3 of the Phase I and up to 50% of the Phase II. Obviously, you have to get through the Phase I before you get to Phase II, where you can subcontract to a company. What companies the SOCOM JATF is looking to partner with in the future, after the feasibility or during the feasibility is done, that's probably not going to be shared because you don't want to go and try to subcontract to them while others will not have that opportunity unless they present it to everyone, full and open, but that would not be fair either. So in the meantime, I would say, do your best in finding the right subcontractor, if you are looking to subcontract. This is an SBIR so you don't need to subcontract, but again up to 1/3 for Phase I, and then whoever gets awarded the Phase I, those are the ones that are going to be the ones you can team up with, and then, agreements will be done accordingly.

6. **Are there any datasets available?**
   Yes, through Meta. We in the JTAF, and in SOCOM in general, we don't collect data sets. That's why we're reaching out to you guys, and we certainly don't have one for Hokkien and that's also why we're reaching out to you guys.
7. **Meta had a real-time Hokkien-English translation before. https://mothership.sg/2022/10/meta-hokkien-translation/ How is this project different? What to improve?**
   I think we stated that upfront, basically we know by all the reports of the Hokkien speakers that it's okay. We don't consider that to be a good enough standard, and it's reliant upon the Cloud, and we very much need this to operate at a high-level disconnect. So those are the things to improve.
8. **Is the DLI metric available?**
   We will try to push that out in the links, but online on YouTube by DLI, some examples of what all the levels look and sound like with a DLI instructor and a native speaker reading at those levels so that you can see and hear what those standards are.
9. **Does the project have any specific preferences for the underlying technology, such as a particular ML model, NLP framework, or software platform?**
   No, we want to leave this open to have proposers' creativity to find the best, most optimal solution, and not come into this with preconceptions that might be incorrect. We're aware of many options out there today. Some of the large language models, basic models from, Google Meta, and other large companies look very promising, but they present challenges when getting them to work at high efficiency on edge devices. Edge being defined as disconnected from the Cloud.
10. **Do you have existing data sets that you can provide to assist with this work? If so, what is known regarding the quantity and quality of the data?**
    At this time, we do not. That's why we're coming to you guys here with this SBIR to help us out with this problem set. We hope anyone out there that's interested in this would absolutely take advantage of the work that Meta has done with AI to build at least a baseline of a data set for this language. That being said, we understand that data sets is one of the largest challenges in this field, and we have looked into acquiring or paying

for new data sets, and that's absolutely something we would consider in the feasibility study to be included. If that is the best, easiest path to a high-quality solution.

11. **What solutions has the organization tried that do not work?**
    Like we've said, we don't really have an existing solution for Hokkien, so there's not many that go in here. We have tried the Meta Hokkien model, and it doesn't work up to our expectations, but that's the only existing capability in this specific area that we're aware of. We spent a couple of years trying to get after high density languages, and every solution that we tried that was just based on a phone, no matter how good the phone, it always locked up, froze up, repeated, or just plain shut down, so we could not find anything that worked at a high level that was phone-only based. It had to be connected to some greater computer. That's what we've learned in our painful experiences. Also we want things to be disconnected, at the edge, but at the same time, being able to have AI integrated into this tech to do updates and learn as we go after these low data language sets.
12. **The focus of the project is on developing new methodology or implementing a solution on portable device?**
    We are looking for a portable solution, and we're going to leave it open to industry for coming up with that solution. We're looking for a template to get after low density languages. The focus on the project is not generating the template. The focus of the project is creating a solution for Hokkien, and if we find that the process of creating that was very effective, and it is extensible to other languages, that'd be a huge plus. But the focus is on getting the capability for Hokkien.
13. **What is the expectation for Phase I deliverables?**
    The Phase I deliverable is a feasibility study. It's a final report that describes: How are you going to solve the problem at that lower technology readiness level (TRL), and how are you going to prototype it? That's really the gist of it. If you want more of what you need to see in the Phase I deliverable feasibility study final report, I think JATF (the technical POC) can kind of jump in and explain what they're looking for. Otherwise, it is in the Phase I, one-page description. And, by the way, whoever gets the Phase I award, the report will be due at the 6-month period. At the 6.5-months mark, which is two weeks after you'll be providing the Phase II proposal. And it'll be based on questions and answers at the 4.5-month period, received from the TPOC team and the PM office. And that will tell us how you're going to prototype it/proof of concept. You'll see this in Contract Data Requirement List (CDRL) #5. That's going to be evaluated before the 7 months Period of Performance (PoP) is completed and then we'll start our Phase II award from that point on.

14. **Are there solutions being used for other languages? What differentiates this language from those?**
    Like we said before, we built up a handful of high-density languages such as Chinese, Russian, Arabic, French, Spanish, Ukrainian, et cetera and we've gotten those to the high

level. We've been working with a company to get there, and it's required a lot of engineering and operator interface so that we can figure out where the machine was failing, so we could finally get it up to that higher level disconnected. And a lot of issues we had to address were like predictive punctuation, continuous conversation, that kind of thing. So those were some of the solutions we found, but again, only with high-density languages, this is our first time into the low-density area. So we're very open to solutions – hardware and software. And that's what differentiates this language from those – primarily data availability. For all the other languages that we've tackled thus far and are currently in-work, they available open-source corpuses. They're much easier to get your own, generate data from, because, there's a lot available for English, Chinese, Russian, French. All of these major languages that have major online presences have much more available and open source corpuses.

15. **Has SOCOM tried to solve the problem through the SBIR program in the past?**
    No, no SBIR program in the entire U.S. government has done anything on Hokkien.
16. **Is there a hardware and software component to this? Does the Phase I prototype involve a hardware component or only a software component?**
    I believe that the SBIR Program Office answered this before, when referencing the deliverables for the Phase I feasibility study. The answer is no, we don't have a hardware or software component, we're expecting you to have something you already use. You are welcome to bring anything to the table. If you find that you don't need custom hardware and software, we are open to using our current form factor for it, which is the phone and the laptop processing connected by a cable. The current solution uses a system, a series of machines, a full end-to-end translation of voice to text translation where the language is first recognized by ASR model, translated in an MT model, and then spoken with a TS model. We expect that Hokkien might be a little more challenging because, from what we understand, there's not a generally accepted method of writing the language down, so that might not be the same for this language. Which is why we want to leave it open for other hardware and software solutions.
    Your report should say what this will look like, whether it's a combination of hardware and software, or just the software. And then it also should go in your proposal, what you're proposing it to be. JATF has said they don't have a preference and they will look at all proposals they receive.
17. **Is there a particular domain that the language model should focus?**
    Yes, conversational speech. The intended use case is an operator talking to a local, on the edge, about day-to-day things. We don't want to focus on any specific military, medical, we just want the best overall conversational speech possible. Our current high-density language device is very specifically a voice-to-voice device. We're operating on the principal that it's very possible that our partner forces might not be literate, so we can't be relying on writing and then reading and then responding. Despite the fact that Hokkien doesn't really have a written form factor.
18. **If we already have a prototype, how does that affect what we would deliver?**

SBIR Program Office: If you think you already have a prototype ready, please continue to propose as a Phase I. It may not be the full technology that we're looking for, it might be something that we can continue to innovate in order to bring it to a full solution that meets the needs of the TPOC team and our requirements. So propose as a Phase I and mention what you've done so far, the solution, and how you're going to bring it to fruition.

19. **Will we get a list of participants so we can match with for partners?**
    Negative. We cannot list the participants, because then everyone who proposes is going to go up to these companies and want to partner, so that would not be fair and open. I would say do your best at what you do best, which is the innovative development of the system, and the partnering part will happen if JATF wants to use SMEs to enable you.
20. **How many phase-I winners will we have?**
    On average, we award three. But we award based on the unique proposals we get.
21. **Will participants be able to work with the company that improved the high density languages to a high level?**
    I can see no reason why anyone who gets awarded a Phase I SBIR wouldn't get partnered up with our other companies, then we can make, ultimately, a better product for the operator.
22. **If we get a phase 1, will it be possible to connect with DLI to get a list of phrases or tests that will allow us to test the solution?**
    First and foremost, a lot of defense language, proficiency tests that are the old ones like the version 4s and 3s are out online so you can get an idea what those standards look like. This is a link to the DLI Site where they have video examples of what each DLPT level looks and sounds like. https://vimeo.com/showcase/139578

    Now I will say a couple more things. DLI's primary mission is to train the force they're not there to push the envelope and experiment, and find new capabilities and whatnot. That's kind of where we come in. We have liaised with DLI for some aspects of this we do have tests that we use when we test the devices to the human metric standard, and we also have a capability being developed to develop our own ratings of anonymous or a random text and video and audio clip, so that we can find out what level it is and use that for testing. So if the question is trying to get after: Can we give you some material to work with the help you test your solution? That short answer is yes. The short answer for getting linked up with DLI, it's certainly possible. But at this point I don't see any reason for that. We can do everything that they would be doing in reference in regard to this effort.
23. **How large should the vocabulary be?**
    When we first started this, everybody wanted to give us a lot of military and medical and engineering technical vocabulary, and that's all great, but it also is very constraining in its scope. Our primary objective is to be able to work with our partner forces to be able to talk to lawyers or farmers equally, and win hearts and minds and help shape and

influence. So our primary focus has been on normal human vocabulary, so that I can talk to them about their farms, their herds, their families, their medical concerns, that type of thing so as large as you think the vocabulary needs to be to win hearts and minds that would be your short answer

A great follow-on for this is why we kind of moved to the DLI method. Those engineering scores are great for comparing one version to another, but they don't do a good enough job in our perception of telling the entire story. One of the ways DLI rates a particular language is the ability to not make any major mistakes that severely impact understanding. We have found that solutions that have very high blue word scores that miss the subject of a sentence will still be completely not understandable. For example, the word cockatoo is a very infrequently used word in multiple languages, but when a translator misses that one word, the entire sentence around it might be correct. The entire science might have an 80+ word errors score, because it only got cockatoo wrong. But now the sentence is completely unintelligible, because it's missed the subject of a sentence, which is kind of why we've moved away from those metrics. It's not to say that you can't use them. That's just not how we give the final tour.

For my pea-brain knuckle dragger foxhole, the device can get all the words wrong and still get the meaning completely wrong. And if that kind of misunderstanding can get operators killed, that creates an issue for us, So that's why we're trying to get after something that is more tangible and measurable on a human metric.

24. **What is the timeline for getting a response to the proposal?**
    SBIR Program Office: We aim for about 30 days or less.
25. **What is the total budget for Phase I?**
    $175,000
26. **Is it possible to fast-track our phase 1, if our solution already meets the expectations?**
    So the proposal will be for seven months, More than likely it's not. It's going to stick to the six months report, and then and then two weeks after the six months for your proposal, and then by the seven month, you know, close outs, et cetera. We'll do our evaluation for Phase II. If you are selected based on your technology, or innovative proposal, and if we go through and there is an opportunity for it that's a discussion to hold after.
27. **What is the commercialization expectation for phase 1?**
    SOCOM specifically focuses on a transition agreement that we have to do between our S&T Director and the funding program office, which means that for us to go from a Phase I to a Phase II, a transition agreement has to be done at the Senior Executive Service (SES) level, signed between a program executive office, someone who has the money to take this into further development, and then S&T Director. That said, that transition agreement has sort of like an acquisition plan, where we state how much money we have, how much we're going to put in in the future, how we're going to work for the contracts Phase III, et cetera. So the first step is to get the transition agreement, once we have that, then we can aware a Phase II. Then we can follow that transition

agreement to say, here is the plan to get it to a Phase III. That's the commercialization part.

Obviously, when you propose you have to provide a commercialization plan to say, "Here's how else I can use this outside of SOCOM, within the International Traffic and Arm Restrictions (ITAR).

28. **Can we focus our propose on one aspect, such as the large langauge model you mentioned?**

    Yes, there's no reason you can't do that. The goal is a complete end-to-end solution and proposals that can achieve that are going to be preferred. But if we have a very strong proposal that has just as a large language model, and we have another really strong proposal that has everything else. We've got no problem in putting them together. But a proposal that does everything, and does everything really well would be preferred.

29. **What IP rights do we have to give away to participate?**

    First off, you keep your IP and data rights as part of the SBIR/STTR program, and that is mandated by Congress. So unless you want to give it up. I've seen a couple of instances where they say they want to give it to SOCOM because they've done all the work they can for whatever they were doing.

    Stakeholders: So the Government, we usually will attain general purpose rights from a defender, and we will not give that away to commercial. But we will share amongst government entities.

    SBIR Program Office: Our Policy Directive says that companies own the IP data rights for this. So we have government purpose rights but there are limitations to that.

    For expected data rights, please, see https://www.acquisition.gov/dfars/252.227-7018-rights-other-commercial-technical-data-and-computer-software%E2%80%94small-business-innovation-research-sbir-program.?searchTerms=sbir%20data%20rights

30. **Will we have access to the current solution, or the dataset that was used in the current solution?**

    The answer to that is no for the Phase I. Like Mike said earlier, if you have a solution that does well, and it works. There's no reason we wouldn't put you together after the fact, our probable end goal after the SBIR, if we have a great solution that works, would be to integrate what the solution that comes out of the end of the SBIR into our current solution, which is another language that can be selected. So with that in mind, preference will be given to solutions that focus on the quality of the actual translation. The components for the actual translation itself, and a little less emphasis will be on user interface, how pretty it looks, how pretty the hardware looks, because that can be sorted out farther down the line, potentially with an integration with our current partner. The data set that was used in the current solution really wouldn't be useful. These are open source corpuses that have other languages that don't include Hokkien, available online for free today. There's no proprietary information that we see that will help you with this process.

At the end of the day, we want you all to succeed and thrive and provide these capabilities to the civilian market. Languages - there's nothing classified about it, so once we can get something for our SOF operators, I'm sure we can get it in front of DoD and civilian sectors pretty quickly, so that's all positives for you all out there trying to help us solve this problem. All we want to be able to do is use what you make for us universally in perpetuity and share amongst Government people. What you guys do with it, and I hope you all get very wealthy off of this, that's great. The data sets – I believe the metadata set is available to anyone and everyone out there. So, we're pointing y'all in that direction. For those who didn't know about it, I would suggest getting a hold of that and using that as your foundation, but we don't have that, so we're not going to be providing it.

31. **For hardware, does it have to be an S22, or can it be an iPhone?**
    We have no hard restrictions against using iPhone, but be aware the government is heavily invested in Android. SOCOM uses ATAK extensively. All of our existing hardware works on Android. So we have a very strong preference for the Android system. If an iPhone system exists and is proposed, it would be great to propose a transition from iPhone to Android, or some other path for it to end up on Android, or any sort of Android. iPhone is not out of the question, we'll find a way.
32. **If our solution performs as well as meta's solution, but works entirely offline, is it possible to work with SOCOM to improve the accuracy?**
    It's a possibility, definitely.